Bayesian Mixtures for Cluster Analysis and Flexible Modeling of Distributions

Dissertation by

Arno Fritsch

submitted to the Department of Statistics,
Technische Universität Dortmund, Germany
in fulfillment of the requirements for
the degree Doktor der Naturwissenschaft

Submitted March 2010

Oral examination held on 11th June 2010

Primary Supervisor: Prof. Dr. Katja Ickstadt

Secondary Supervisor: Prof. Dr. Claus Weihs

Acknowledgements

Thanks to ...

- ... my supervisor Katja Ickstadt, who gave me enough time to pursue my own research and always encouraged me in my work.
- ... to Claus Weihs for refereeing the thesis and to Jörg Rahnenführer and Marco Grzegorczyk for completing the examination committee.
- ... the European Union and the state of North Rhine-Westphalia for paying me money during my time at the Centre of Applied Proteomics (ZAP).
- ... my colleagues from the chair "Mathematische Statistik und biometrische Anwendungen": Björn Bornkamp, Evgenia Freis, Martin Schäfer and Jakob Wieczorek, for proofreading this thesis. You earned your candy bars well!
- ... Brigitte Koths, Eva Brune and Jadwiga Schall for administrative support.
- ... Uwe Ligges and his "Rechnerhiwis" Sebastian Krey and Olaf Mersmann for making my computer run almost all of the time.
- ... the people from the seventh floor for a nice working environment and the talks during lunch and the weekly "cake break".
- ... Oliver Kuß for the idea and data for the goalkeeper application.
- ... Björn Bornkamp for many interesting discussions about Bayesian statistics, for providing me with useful hints and references and for help with programming in C.

- ... my family for supporting me with the non-statistical aspects of life.
- ... my sister Barbara for giving me the idea to study statistics.
- ... my daughter Zoe for being a well-behaved baby and allowing me to get enough sleep to finish my thesis in time.
- ... my wife Anna for being my reference prior.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Classical statistical methods often make parametric assumptions about distributions, the most common one being the assumption of a normal distribution. Although parametric distributions provide a reasonable approximation to the truth in many cases, there are also many situations were their use is not justified. A wide variety of nonparametric methods have been developed to alleviate this problem. These are, for example, based on ranks or the empirical distribution function. Many nonparametric methods, however, lack interpretability. Finite mixture models assume that a distribution is a combination of several parametric distributions. They offer a compromise between the interpretability of parametric models and the flexibility of nonparametric models. Although they are strictly speaking still parametric models, they are therefore sometimes considered to be semiparametric. Figure 1.1 shows some examples of mixtures of two normals that show features that normal distributions cannot possess, i.e., bimodality, heavy tails and skewness. While the densities shown have only two components and are still relatively close to the normal distribution, Marron and Wand (1992) give a collection of mixtures of normals that have densities with a much wider range of shapes.

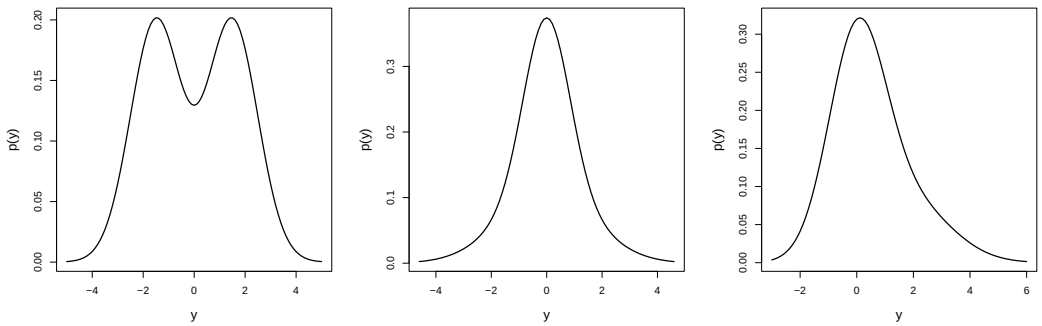

Figure 1.1: Densities of three mixtures of two normals exhibiting bimodality, heavy tails and skewness. The densities belong to mixture distributions given by $0.5 \cdot N(-1.5, 1) + 0.5 \cdot N(1.5, 1)$, $0.549 \cdot N(0, 0.676) + 0.451 \cdot N(0, 2.799)$ and $0.7 \cdot N(0, 1) + 0.3 \cdot N(1.75, 2.25)$ ($N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2). The second mixture is chosen to minimize the absolute distance to the density of the t-distribution with 4 degrees of freedom, integrated between the 0.01 and 0.99 quantiles of the t-distribution.

Fitting finite mixture models in a Bayesian framework offers many advantages. As will be discussed in Chapter 2 it allows valid inference also for small samples and, using only weak prior information, can solve problems with unbounded likelihood functions that can, for example, occur with finite mixtures of normals. The Bayesian approach also offers an elegant extension of finite to (countable) infinite mixture models. These can be shown to provide the same flexibility as nonparametric models, as the posterior on the space of distributions can be shown to accumulate around the true distribution. These results are reported in Chapter 3 where, as an example, Bayesian mixtures are also used to flexibly model the goalkeeper's effect on the probability of saving a penalty.

Depending on the application the components of mixture models can either be viewed as just a means to the flexible modeling of a distribution or as defining subgroups of a population with different parametric distributions. In the latter case mixture models can be used for cluster analysis and the group structure of a sample of observations can be inferred. In a cluster analysis application Bayesian mixtures, besides the already mentioned advantages, allow the estimation of the number of clusters at the same time as the cluster-specific parameters.

A drawback is, however, that the standard approach for fitting Bayesian models, Markov Chain Monte Carlo (MCMC), unfortunately leads to inferential difficulties in this case. Since the likelihood (and usually the posterior as well) of a mixture model is invariant to a permutation of component labels, the labels associated with the clusters can change during the MCMC run, a phenomenon called label-switching. Label switching has to be addressed appropriately before clustering inference can be drawn. The problem gets severe, if the number of clusters is indeed allowed to vary during the MCMC run. Existing methods to deal with label-switching and a varying number of components are reviewed in Chapter 5 and new approaches are proposed for both situations. One of the new approaches, the maximization of the posterior expected adjusted Rand index, makes use of similarity measures for clusterings. These are therefore reviewed in Chapter 4, along with a short general introduction to cluster analysis and a description of the relation of finite mixture models and common clustering criteria. Finally, the new approaches are compared to the previous methods on simulated and real data. The real data used for cluster analysis are two gene expression data sets and Fisher's iris data.

Chapter 2

Bayesian Mixtures

2.1 Finite Mixtures

2.1.1 Basic Definition

Finite mixture distributions have a long history in statistics dating back to the work of Pearson (1894) who first fitted a mixture with two normal components to a data set. Only a brief introduction to the theory of finite mixture models can be given here, which focuses on Bayesian approaches. More detailed discussion of many aspects sketched here can be found in Frühwirth-Schnatter (2006). For more details on the frequentist approach to finite mixture modeling see McLachlan and Peel (2000).

A random variable or vector Y that takes values in $\mathcal{Y} \subseteq \mathbb{R}^p$ has a finite mixture distribution if its probability density function (existence of densities with respect to the Lebesgue measure will be assumed throughout this thesis) can be written as

$$p(y|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(y|\theta_k) \ , \tag{2.1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ and $p(y|\theta_k)$ is a probability

density with parameter θ_k and π_k is the weight of the kth component. The weights are restricted by

$$\pi_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1 \ ,$$

so that $\boldsymbol{\pi}$ is a point on the K-dimensional simplex. While the component densities $p(y|\theta_k)$ can be arbitrary parametric densities, they are often taken to be members of the exponential family. In this thesis $p(y|\theta_k)$ will in most cases be univariate or multivariate normal distributions which will then be denoted by $\phi(y|\mu_k, \sigma_k^2)$ and $\phi(y|\mu_k, \Sigma_k)$, so that $\theta_k = (\mu_k, \sigma_k^2)'$ and $\theta_k = (\mu_k, \Sigma_k)'$. For now the number of components $K \in \mathbb{N}$ will be assumed to be known.

An equivalent formulation of model (2.1) is obtained by introducing an allocation variable Z where

$$p(y|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} P(Z = k|\boldsymbol{\pi}) p(y|Z = k, \boldsymbol{\theta})$$

with

$$\begin{aligned} P(Z = k|\boldsymbol{\pi}) &= \pi_k \quad \text{and} \\ p(y|Z = k, \boldsymbol{\theta}) &= p(y|\theta_k) \ . \end{aligned}$$

This formulation is needed for cluster analysis, as $Z = k$ denotes membership in the kth cluster in this case. If $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are known a simple application of Bayes' theorem gives the distribution of $Z|y$ as

$$P(Z = j|y, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{\pi_j p(y|\theta_j)}{\sum_{k=1}^{K} \pi_k p(y|\theta_k)} \ . \tag{2.2}$$

Estimating Z to be equal to the value of j that maximizes (2.2) is often referred to as the Bayes classifier for Z.

If a sample of independent random variables $Y_1, \ldots, Y_n$ from (2.1) is observed, yielding observations $\mathbf{y} = (y_1, \ldots, y_n)'$, the likelihood function is given

by

$$p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\pi}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta},\boldsymbol{\pi}) = \prod_{i=1}^{n}\left(\sum_{k=1}^{K}\pi_k p(y_i|\theta_k)\right) \quad . \tag{2.3}$$

This can again be written in an alternative form with a vector $\mathbf{Z} = (Z_1,\ldots,Z_n)'$ of allocation variables via

$$p(\mathbf{y}|\mathbf{Z},\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{k=1}^{K} p(y_i|\theta_k)^{I(Z_i=k)} \tag{2.4}$$

and

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{I(Z_i=k)} = \prod_{k=1}^{K}\pi_k^{\sum_{i=1}^{n} I(Z_i=k)} \quad , \tag{2.5}$$

where $I(.)$ denotes an indicator function. In the following $\sum_{i=1}^{n} I(Z_i = k)$, the number of observations associated with component k, will be abbreviated by n_k. The likelihood (2.3) is recovered by summing $p(\mathbf{y}|\mathbf{Z},\boldsymbol{\theta})p(\mathbf{Z}|\boldsymbol{\pi})$ over the K^n possible values of $\mathbf{Z}$.

2.1.2 Priors

For Bayesian inference a prior distribution has to be assigned to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, which are usually assumed to be independent a priori, so that

$$p(\boldsymbol{\theta},\boldsymbol{\pi}) = p(\boldsymbol{\theta})p(\boldsymbol{\pi}) \quad .$$

For the prior on the weights $\boldsymbol{\pi}$ a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)'$ will be taken, which will be denoted by $Dir(\boldsymbol{\alpha})$. The density is given by

$$p(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1} \propto \prod_{k=1}^{K}\pi_k^{\alpha_k-1} \quad .$$

The Dirichlet distribution is the multivariate extension of the Beta distribution. The reason for using it is that it is the conjugate prior for $\boldsymbol{\pi}$ in (2.5),

because

$$p(\boldsymbol{\pi}|\mathbf{Z}) \propto p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k+n_k-1} \ ,$$

which is a $Dir(\alpha_1+n_1,\ldots,\alpha_K+n_K)$. This property will turn out to be useful for fitting the model (2.3).

The specific prior assigned to $\boldsymbol{\theta}$ depends of course on the form of the $p(y|\theta_k)$. Roeder and Wasserman (1997) showed that in general assigning an improper prior to $\boldsymbol{\theta}$ leads to an improper posterior. This is due to the fact that similar to the likelihood (2.3) the posterior $p(\boldsymbol{\theta},\boldsymbol{\pi}|\mathbf{y})$ can be written as a summation over all possible allocations of $\mathbf{Z}$. For values of $\mathbf{Z}$ where at least one component is empty the observations provide no information on some of the θ_k's. For these $\mathbf{Z}$ an improper prior leads to $p(\boldsymbol{\theta},\boldsymbol{\pi}|\mathbf{y},\mathbf{Z})$ being improper which then leads to the whole posterior being improper. We will therefore tend to assign relative vague but proper prior distributions to $\boldsymbol{\theta}$. If the component densities are from the exponential family it is again convenient for model fitting to use conjugate priors. As an example we will consider the case where the component densities are univariate normals. A prior suggested, for example, by Bensmail et al. (1997) is the conditionally conjugate prior

$$\begin{aligned} p(\boldsymbol{\theta}) &= \prod_{k=1}^{K} p(\mu_k|\sigma_k^2)p(\sigma_k^2) \\ \mu_k|\sigma_k^2 &\sim N(b_0,\sigma_k^2/v) \\ \sigma_k^{-2} &\sim Ga(c_0,C_0) \ . \end{aligned} \tag{2.6}$$

The component parameters $\theta_k = (\mu_k,\sigma_k^2)'$ are assumed to be independent, whereas μ_k and σ_k^2 are dependent in each component. $Ga(c_0,C_0)$ denotes a Gamma distribution with expectation c_0/C_0. The posterior distribution $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})$ for a known allocation vector $\mathbf{Z}$ is then available in closed form,

which is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z}) &= \prod_{k=1}^{K} p(\mu_k|\sigma_k^2,\mathbf{y},\mathbf{Z})p(\sigma_k^2|\mathbf{y},\mathbf{Z}) \\ \mu_k|\sigma_k^2,\mathbf{y},\mathbf{Z} &\sim N\left(\frac{v}{v+n_k}b_0 + \frac{n_k}{v+n_k}\bar{y}_k, \frac{\sigma_k^2}{v+n_k}\right) \\ \sigma_k^{-2}|\mathbf{y},\mathbf{Z} &\sim Ga\left(c_0+\frac{n_k}{2}, C_0+\frac{1}{2}\left[n_k s_{y,k}^2 + \frac{v\,n_k}{v+n_k}(\bar{y}_k-b_0)^2\right]\right) \quad . \end{aligned}$$

where $\bar{y}_k$ and $s_{y,k}^2$ are the sample mean and variance of the observations with $Z_i = k$. Bensmail et al. (1997) choose the hyperparameters to be data-dependent as $b_0 = \bar{y}$, $v = 1$, $c_0 = 2.5$ and $C_0 = 0.5 \cdot s_y^2$, where $\bar{y}$ and s_y^2 are the sample mean and variance of all observations.

2.1.3 Identifiability

A problem with mixture models is that they are not identifiable, i.e., there are distinct parameter values $(\boldsymbol{\theta},\boldsymbol{\pi})'$ and $(\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)'$ so that

$$p(y|\boldsymbol{\theta},\boldsymbol{\pi}) = p(y|\boldsymbol{\theta}^*,\boldsymbol{\pi}^*) \quad \text{for almost all } y \in \mathcal{Y} \ .$$

One reason for this is that a permutation of the component labels does not change $p(y|\boldsymbol{\theta},\boldsymbol{\pi})$, as just the order of summation in (2.1) is changed. So for any permutation ν (i.e., ν being a bijective map from $\{1,\ldots,K\}$ into $\{1,\ldots,K\}$) it holds that $p(y|\boldsymbol{\theta},\boldsymbol{\pi}) = p(y|\nu(\boldsymbol{\theta}),\nu(\boldsymbol{\pi}))$. This leads to the likelihood (2.3) having $K!$ identical modes. The same is true for the posterior distribution $p(\boldsymbol{\theta},\boldsymbol{\pi}|\mathbf{y})$ if the priors $p(\theta_k)$ and $p(\pi_k)$ are identical for all k, which will usually be the case. This will lead to problems for the Bayesian estimation of $(\theta_k,\pi_k)'$ or Z, which will be considered in detail in Chapter 5. Note, however, that identifiability is not an issue if one wants to estimate a function of the parameters that is invariant to a permutation of component labels, for example, the mixture density $p(y|\boldsymbol{\theta},\boldsymbol{\pi})$ itself.

Unidentifiability of $p(y|\boldsymbol{\theta}, \boldsymbol{\pi})$ also arises when a $(K+1)$th component is added and either $\pi^*_{K+1} = 0$ or $\theta^*_{K+1} = \theta_k$ and $\pi^*_{K+1} + \pi^*_k = \pi_k$. As will be discussed in Subsection 2.1.5 at least the situation where $\pi^*_{K+1} = 0$ can be effectively dealt with by choosing an appropriate prior for the weights $\boldsymbol{\pi}$.

While the above mentioned problems with identifiability arise for component densities $p(y|\theta_k)$ of any family of distributions, for some families the additional problem of generic identifiability arises. One such family is the family of uniform distributions where, for example,

$$\begin{aligned} Y_1 &\sim \frac{1}{2}U[-2,1] + \frac{1}{2}U[-1,2] \quad \text{and} \\ Y_2 &\sim \frac{1}{3}U[-1,1] + \frac{2}{3}U[-2,2] \end{aligned}$$

have the same density. Another example is a mixture of Bernoulli distributions $\sum_{k=1}^{K} \pi_k Bern(p_k)$, which turns out to have the same density as a single Bernoulli with parameter $\sum_{k=1}^{K} \pi_k p_k$. Teicher (1963) proved that, among others, univariate mixtures of normals and Gamma distributions are generically identifiable. These results were extended to multivariate mixtures of normals by Yakowitz and Spragins (1968).

2.1.4 Model Fitting

Modern computational approaches for fitting the finite mixture model make use of the representation of the likelihood given by (2.4) and (2.5), which includes the allocation vector $\mathbf{Z}$. In a frequentist setting maximum likelihood estimates are usually obtained via the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This algorithm iterates between the E-step, which for finite mixture models consists of computing the expectations

$$E(I(Z_i = k)|y_i, \boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}) = P(Z_i = k|y_i, \boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}) \doteq \widehat{Z}_{ik}^{(t+1)}$$

given the values $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\pi}^{(t)}$ at iteration t, where the '$\doteqdot$'-symbol denotes a definition. The $\widehat{Z}_{ik}^{(t+1)}$ are then given by (2.2).

In the M-step the expectation of the logarithm of the likelihood

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\widehat{Z}_{ik}^{(t+1)}(\log\pi_k + \log p(y_i|\theta_k)) \tag{2.7}$$

is maximized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ to obtain $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\pi}^{(t+1)}$. The value of $\boldsymbol{\pi}$ that maximizes (2.7) is in general given by

$$\pi_k^{(t+1)} = \sum_{i=1}^{n}\frac{\widehat{Z}_{ik}}{n} \ , \tag{2.8}$$

whereas the value of $\boldsymbol{\theta}$ that maximizes (2.7) depends of course on the distribution family of $p(y|\theta_k)$. It is often available in closed form.

Dempster et al. (1977) show that an iteration of the EM algorithm does not decrease the likelihood $p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\pi})$, so that the algorithm is guaranteed to converge if the likelihood is bounded from above. The EM algorithm can also be applied in a Bayesian context, where the aim is to find values of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ that maximize the posterior density $p(\boldsymbol{\theta},\boldsymbol{\pi}|\mathbf{y})$, also known as the MAP estimate. This approach is, for example, taken by Fraley and Raftery (2007). The E-step is unchanged and for the M-step the terms $\log p(\boldsymbol{\pi})$ and $\log p(\boldsymbol{\theta})$ representing the priors are included in (2.7). For example, under the $Dir(\boldsymbol{\alpha})$ prior for $\boldsymbol{\pi}$ (2.8) changes to

$$\pi_k^{(t+1)} = \sum_{i=1}^{n}\frac{\widehat{Z}_{ik}+\alpha_k-1}{n+\sum_{k=1}^{K}\alpha_k-K} \ .$$

Alternatively, Markov Chain Monte Carlo (MCMC) methods can be applied to obtain a sample of the posterior distribution. These methods construct a Markov chain that has the posterior as its stationary distribution. If the chain is run for some time the draws will come from the posterior distribution. More details on MCMC methods can, for example, be found in Robert

and Casella (2004). For finite mixtures a special MCMC method, the Gibbs sampler (Gelfand and Smith, 1990), is usually employed. The Gibbs sampler starts from some allocation $\mathbf{Z}^{(0)}$ and iterates the following steps, where $\mathbf{y}_k^{(t)} = \{y_i : Z_i^{(t)} = k\}$:

1a) $\boldsymbol{\pi}^{(t+1)}|\mathbf{Z}^{(t)}$ is sampled from $Dir(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.

1b) For $k = 1, \ldots, K$: $\theta_k^{(t+1)}|\mathbf{y}, \mathbf{Z}^{(t)}$ is sampled from $p(\theta_k|\mathbf{y}_k^{(t)})$.

2) For $i = 1, \ldots, n$: $Z_i^{(t+1)}|y_i, \boldsymbol{\theta}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}$ is sampled from the multinomial distribution given by (2.2).

If a conjugate prior $p(\boldsymbol{\theta})$ has been chosen as in (2.6) the posterior $p(\theta_k|\mathbf{y}_k)$ is available in closed form. The first B iterations, which the chain needs to reach its stationary distribution, are discarded and inferences are based on the remaining sample $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)})$, $t = B + 1, \ldots$. Optionally this sample can be thinned to reduce autocorrelation by keeping only every jth draw. The Gibbs sampler can be seen as a stochastic version of the EM algorithm.

As pointed out by Chen and Liu (1996) an alternative Gibbs sampler is available that consists of sampling from the posterior of the allocations

$$p(\mathbf{Z}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{Z})p(\mathbf{Z})$$

directly. For this the distributions

$$\begin{aligned} p(\mathbf{y}|\mathbf{Z}) &= \prod_{k=1}^{K} \int p(\mathbf{y}_k|\theta_k)p(\theta_k)d\theta_k \quad \text{and} \\ p(\mathbf{Z}) &= \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})d\boldsymbol{\pi} \end{aligned}$$

must be available in closed form, which for the former is the case if $p(\theta_k)$ is conjugate. For a $Dir(\boldsymbol{\alpha})$ prior on $\boldsymbol{\pi}$ with a symmetric parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)'$ it turns out that

$$p(\mathbf{Z}) = \frac{\Gamma(K\alpha) \prod_{k=1}^{K} \Gamma(\alpha + n_k)}{\Gamma(K\alpha + n)\Gamma(\alpha)^K} \quad . \tag{2.9}$$

The distribution of $Z_i|\mathbf{Z}_{-i}, \mathbf{y}$ is then used for Gibbs sampling, where $\mathbf{Z}_{-i} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)'$. Details can be found in Chen and Liu (1996) or Frühwirth-Schnatter (2006, Chapter 3.4).

2.1.5 Advantages of a Bayesian Approach

So why be Bayesian when fitting a finite mixture model? It is usually not the case that substantive prior information is included in either $p(\boldsymbol{\theta})$ or $p(\boldsymbol{\pi})$. However, to have weakly informative priors on these parameters as in (2.6) can be of advantage. It is well known that the EM algorithm for ML estimation of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ often diverges to a point of infinite likelihood. For many mixtures, e.g., univariate and multivariate mixtures of normals where the components have different variances, the likelihood is unbounded. Consider a mixture of univariate normals. If $\mu_k = y_i$ for some k and i the likelihood (2.3) contains a factor that is proportional to $1/\sigma_k$. If $\sigma_k^2 \to 0$ the likelihood will tend to infinity. It also often happens that the EM algorithm converges to so-called spurious modes of the likelihood, that correspond to a component of small variance being fitted to a small group of close observations. McLachlan and Peel (2000, Chapter 3.10) give a detailed discussion on spurious modes. A possible solution due to Hathaway (1985) is to impose the constraint $\min_{k,j} \sigma_k^2/\sigma_j^2 \geq c > 0$. The problem is, however, to find a value of c that is large enough to get rid of most spurious modes but not so large that the constraint is not fulfilled by the true parameter value (Hathaway, 1985). If a $Ga(c_0, C_0)$ prior for σ^{-2} is used as in (2.6), Frühwirth-Schnatter (2006, p. 180) shows that σ_k^2/σ_j^2 follows an F-distribution with parameters $2c_0$ and $2c_0$. For $c_0 > 2$ this density has finite variance and is bounded away from 0, leading to a stochastic version of Hathaway's constraint. A stochastic constraint is more flexible than a strict boundary, instead of completely ruling

out some solutions they are only downweighted by the prior. The model is thus less sensitive to the choice of c_0 than of c.

Often it is desirable to have all $\pi_k > 0$, so that the model can be distinguished from a $(K-1)$-dimensional mixture. The Dirichlet distribution on $\boldsymbol{\pi}$ is usually chosen to be symmetric in the components, an exception being outlier modeling, see e.g., Verdinelli and Wasserman (1991), so that $\alpha_1 = \ldots = \alpha_K \doteqdot \alpha$. Then marginally $\pi_k \sim Beta(\alpha, (K-1)\alpha)$, which can be bounded away from 0 by choosing $\alpha > 1$, Frühwirth-Schnatter (2006, pp. 104-105), for example, advocates using $\alpha = 4$.

So even with relatively vague priors it is possible to rule out undesirable behavior of the parameters. Another advantage is that for finite mixture models the sample size has to be fairly large before the asymptotic theory of maximum likelihood applies (McLachlan and Peel, 2000, p. 68). A Bayesian approach via Gibbs sampling draws directly from the posterior distribution and allows valid inference also for small samples.

2.1.6 Choice of Number of Components K

Up till now the number of components K has been assumed to be known. If K is unknown one is faced with a model selection problem. If the finite mixture model with K components is denoted by $\mathcal{M}_K$, one has to choose one of the models $\mathcal{M}_1, \ldots, \mathcal{M}_{K_{max}}$. Often standard model selection criteria of the form

$$-2 \log p(\mathbf{y}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, \mathcal{M}_K) + \lambda \cdot d_K$$

are minimized, where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\pi}}$ refer to either ML or MAP estimates, λ is a penalty parameter for model complexity and d_K is the number of parameters of $\mathcal{M}_K$. Choosing $\lambda = 2$ or $\lambda = \log(n)$ leads to the well-known information criteria AIC (Akaike, 1974) or BIC (Schwarz, 1978), respectively. It should be

noted that because of the identifiability issues discussed in Subsection 2.1.3 the regularity conditions for the asymptotic justification of these criteria do not hold. Nevertheless the criteria are often employed for finite mixtures. For AIC it has been found by many authors that the true number of components is overestimated, see e.g., McLachlan and Peel (2000, pp. 217-220). Use of BIC is advocated among others by Roeder and Wasserman (1997) and Fraley and Raftery (2002) and it selects the correct number of components if the component densities $p(y|\theta_k)$ are correctly specified.

The BIC gives an asymptotic approximation to the logarithm of the marginal likelihood

$$p(\mathbf{y}|\mathcal{M}_K) = \int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}, \mathcal{M}_K) p(\boldsymbol{\theta}, \boldsymbol{\pi}|\mathcal{M}_K) d\boldsymbol{\theta} d\boldsymbol{\pi} \ ,$$

see Kass and Raftery (1995) for details. Given an MCMC sample $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)})$ the integral can also be approximated using numerical methods which are discussed in detail in Frühwirth-Schnatter (2006, pp. 139-165). K can then be chosen so that the marginal likelihood is maximized.

It is also possible to treat $\mathcal{M}_K$ as a random variable, to assign it a prior distribution and to consider the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\pi}|\mathcal{M}_K, \mathbf{y})p(\mathcal{M}_K|\mathbf{y})$. To sample from such a posterior an MCMC sampler that includes moves that change the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ is needed. Such a sampler is given by the reversible jump algorithm of Green (1995), which has been applied to finite mixture models by Richardson and Green (1997). They consider moves that split a component in two or merge two components as well as the birth or death of an empty component. Care must be taken to define these dimension-changing moves in such a way that they can be reverted. See Frühwirth-Schnatter (2006, pp. 129-139) for an introduction to reversible jump MCMC for finite mixture models.

Although reversible jump MCMC has been applied successfully to a vari-

ety of mixture models, in this thesis we will focus on another approach, that will be considered in the next section. The approach extends the finite mixture model to an infinite mixture, which implicitly places a prior distribution on the number of components K.

2.2 Infinite Mixtures

In this section it is described how infinite mixture models can be defined directly via a stick-breaking construction as well as indirectly via discrete random mixing measures. A focus will be on mixtures arising from the Dirichlet process (Ferguson, 1973), as they are historically the first models of this kind and still the most commonly applied.

2.2.1 Stick-Breaking Priors

It is possible to extend the finite mixture model of Section 2.1 to a (countable) infinite mixture model of the form

$$p(y_i|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{\infty} \pi_k p(y_i|\theta_k) \ . \tag{2.10}$$

As in the finite mixture case the prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are assumed independent, where the distribution of the π_k needs to fulfill the conditions $\pi_k \geq 0$ and

$$\sum_{k=1}^{\infty} \pi_k = 1 \quad \text{almost surely} \ . \tag{2.11}$$

Ishwaran and James (2001) define $p(\boldsymbol{\pi})$ by a so-called stick-breaking construction, where

$$\begin{aligned} \pi_1 &= V_1 \quad \text{and} \\ \pi_k &= (1 - V_1)(1 - V_2) \dots (1 - V_{k-1}) V_k \ , \quad k \geq 2 \ , \end{aligned} \tag{2.12}$$

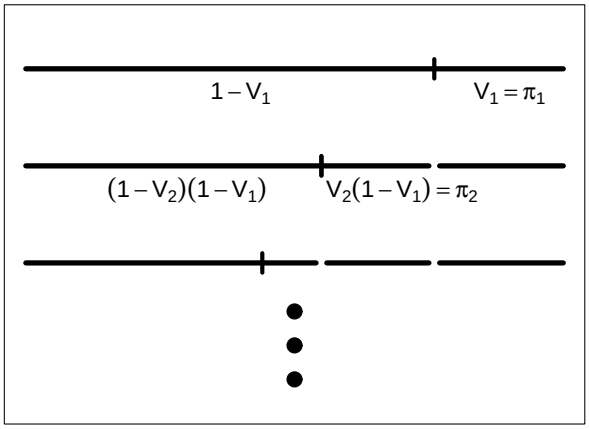

Figure 2.1: Illustration of stick-breaking construction process.

and the V_k are independently $Beta(a_k, b_k)$ distributed with $a_k, b_k > 0$. Starting with a stick of length 1, pieces are successively broken off. The proportion of the remaining stick that is broken off is determined by V_k. Figure 2.1 illustrates this construction. Ishwaran and James (2001) give conditions on the a_k and b_k for (2.11) to hold, it is for example fulfilled iff $\sum_{k=1}^{\infty} \log(1 + \frac{a_k}{b_k}) = \infty$. An example where this is not the case is given by $a_k = a$ and $b_k = k^2$, since $\log(1 + \frac{a}{k^2}) < \frac{a}{k^2}$ and $\sum_{k=1}^{\infty} \frac{a}{k^2} < \infty$.

The most common stick-breaking priors arise from the Dirichlet process of Ferguson (1973) and from the Pitman-Yor process (Pitman and Yor, 1997). In these cases the π_k's have stick-breaking constructions that fulfill condition (2.11) with $a_k = 1$ and $b_k = \alpha > 0$ for the Dirichlet process and $a_k = 1 - \beta$ and $b_k = \alpha + k\beta$ for the Pitman-Yor process, where $0 \leq \beta < 1$ and $\alpha > -\beta$. In the next subsection the Dirichlet process will be considered in more detail as the applications in Section 3.2 and 5.5 make use of it.

2.2.2 The Dirichlet Process

An alternative way to express a mixture model is as

$$p(y_i|\boldsymbol{\theta}, \boldsymbol{\pi}) = \int p(y_i|\varphi) dG(\varphi) \ ,$$

where G is a discrete probability measure with atoms $\boldsymbol{\theta}$ and weights $\boldsymbol{\pi}$. If $(\boldsymbol{\theta}, \boldsymbol{\pi})$ are assigned a prior, G becomes random. For infinite mixtures it is common to define a prior directly for the measure G. In this case the model can be written as

$$\begin{aligned} y_i|\varphi_i &\sim F(\varphi_i) \\ \varphi_i|G &\sim G \\ G &\sim D \ , \end{aligned} \tag{2.13}$$

where $F(\varphi)$ is the distribution with density $p(y|\varphi)$, and D specifies a distribution over the distributions G. The most common choice for D is the Dirichlet process (DP). Ferguson (1973) introduced the DP as a random probability measure on a measurable space $(\Omega, \mathcal{A})$ with mass parameter α and base measure G_0, written $G \sim DP(\alpha, G_0)$. Its defining property is that for every finite partition $(A_1, \ldots, A_l)$ of Ω,

$$(G(A_1), \ldots, G(A_l)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_l)) \ . \tag{2.14}$$

Ferguson shows that the distributions (2.14) are consistent over all partitions and thus define a stochastic process. Since (A, A^C) is a partition it follows that $G(A)$ has a $Beta(\alpha G_0(A), \alpha[1 - G_0(A)])$ distribution for all $A \in \mathcal{A}$ and thus

$$E(G(A)) = G_0(A) \qquad \text{and} \tag{2.15}$$

$$Var(G(A)) = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1} \ . \tag{2.16}$$

The base measure G_0 can therefore be seen as the expectation of G, whereas α determines its variability around G_0. Ferguson (1973) also gives the covariance of G for any two sets $A_1, A_2 \in \mathcal{A}$ as

$$Cov(G(A_1), G(A_2)) = \frac{G_0(A_1 \cap A_2) - G_0(A_1)G_0(A_2)}{\alpha + 1} \quad .$$

Sethuraman (1994) showed that the DP can be written as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad , \tag{2.17}$$

where the θ_k are drawn from G_0 independently from each other and the π_k's follow the stick-breaking construction given in the last subsection. Because of the discreteness of the DP, exhibited in (2.17), it is generally not used to model observations directly. Rather an underlying parameter φ is assumed to follow a DP, as in (2.13). Such Dirichlet process mixture models were first considered by Antoniak (1974). In the following we will therefore assume that the Dirichlet process is used to model the distribution of a parameter φ.

The Dirichlet process also has a conjugacy property. If $\varphi_1, \ldots, \varphi_n | G \overset{i.i.d.}{\sim} G$ and $G \sim DP(\alpha, G_0)$ then the posterior distribution $G|\varphi_1, \ldots, \varphi_n$ is again a Dirichlet process with parameters $\alpha + n$ and

$$G_0^* = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\varphi_i} \quad . \tag{2.18}$$

The distribution function of G_0^* is thus a weighted mean of the distribution function of G_0 and the empirical distribution function of the φ_i's.

Blackwell and MacQueen (1973) showed that the predictive distribution of $\varphi_{n+1}|\varphi_1, \ldots, \varphi_n$, where G has been integrated out, is given by (2.18). There is thus positive probability that φ_{n+1} is equal to one of the previously drawn φ's. The successive distributions of $\varphi_i|\varphi_1, \ldots, \varphi_{i-1}$ have been described by Pitman (mentioned in Aldous (1985)) with the metaphor of the Chinese

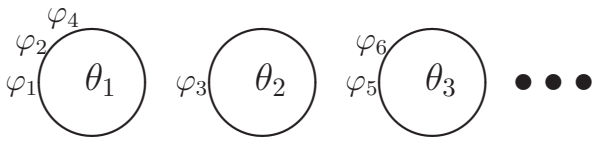

Figure 2.2: Chinese restaurant process.

restaurant process, where customers $\varphi_1, \varphi_2, \ldots$ arrive in a restaurant with infinitely many tables, where a specific dish θ_k is served at each table. If a newly arriving guest knows someone already sitting at a table he joins that table, otherwise he starts a new table, see Figure 2.2 for an illustration. A related metaphor is that of a Pólya urn. Let $\theta_1, \ldots, \theta_K$ be the unique values of $\varphi_1, \ldots, \varphi_n$ with $K \leq n$ and, as in the finite mixture model, introduce allocation variables Z_i such that $\varphi_i = \theta_{Z_i}$. The distribution of $Z_i|\mathbf{Z}_{i-1}$ with $\mathbf{Z}_{i-1} = (Z_1, \ldots, Z_{i-1})'$ is then given by

$$P(Z_i = k|\mathbf{Z}_{i-1}) = \begin{cases} \frac{n_{k,i-1}}{\alpha+i-1} & , \quad k = 1, \ldots, K \\ \frac{\alpha}{\alpha+i-1} & , \quad k = K+1 \end{cases} , \tag{2.19}$$

with $n_{k,i-1}$ being the number of $\mathbf{Z}_{i-1}$ equal to k. The successive draws of $Z_i|\mathbf{Z}_{i-1}$ thus follow an urn scheme, where, if a ball of a specific color is drawn, the ball is replaced and another one of the same color is added and with probability $\alpha/(\alpha + i - 1)$ a ball of a new color is drawn.

Neal (2000) and Green and Richardson (2001) considered the derivation of the Dirichlet process as limit of a finite mixture model. For a finite mixture model with K^* components, where a priori $\theta_k \overset{i.i.d}{\sim} G_0$ and

$\boldsymbol{\pi} \sim Dir(\alpha/K^*, \ldots, \alpha/K^*)$, it can be obtained that

$$\begin{aligned} P(Z_i = k|\mathbf{Z}_{i-1}) &= \frac{P(Z_i = k, \mathbf{Z}_{i-1})}{P(\mathbf{Z}_{i-1})} \\ &\overset{(2.9)}{=} \frac{\Gamma(\frac{\alpha}{K^*} + n_{k,i-1} + 1)}{\Gamma(\alpha + i)} \frac{\Gamma(\alpha + i - 1)}{\Gamma(\frac{\alpha}{K^*} + n_{k,i-1})} \\ &= \frac{\frac{\alpha}{K^*} + n_{k,i-1}}{\alpha + i - 1} \ , \end{aligned}$$

which for $K^* \to \infty$ leads to

$$P(Z_i = k|\mathbf{Z}_{i-1}) = \frac{n_{k,i-1}}{\alpha + i - 1} \quad \text{and} \tag{2.20}$$

$$P(Z_i \neq k \, \forall k \leq K|\mathbf{Z}_{i-1}) = \frac{\alpha}{\alpha + i - 1} \ . \tag{2.21}$$

In (2.21) Z_i can be set to $K + 1$. The distribution of $Z_i|\mathbf{Z}_{i-1}$ given by (2.20) and (2.21) is then equal to (2.19), so that the defined models are the same.

2.2.3 Model Fitting

In this subsection two Gibbs sampling algorithms for fitting infinite mixture models of the form (2.13) are introduced. They are described for the special case of Dirichlet process mixture models but can be used in a similar fashion for other stick-breaking priors (Ishwaran and James, 2001). The first algorithm relies on the Pólya urn scheme (2.19). Given a sample $y_1, \ldots, y_n$ each Z_i is taken in turn to be the last observation arising from the urn, which is valid because the Z_i are exchangeable. Then

$$P(Z_i = k|\mathbf{Z}_{-i}, \mathbf{y}, \boldsymbol{\theta}) \propto \begin{cases} \frac{n_{k,-i}}{\alpha+n-1} \, p(y_i|\theta_k) & , \quad k = 1, \ldots, K \\ \frac{\alpha}{\alpha+n-1} \int p(y_i|\theta) dG_0(\theta) & , \quad k = K+1 \end{cases} , \tag{2.22}$$

with $n_{k,-i}$ being the number of $\mathbf{Z}_{-i}$ equal to k. The Pólya urn Gibbs sampler repeats the following steps starting from an allocation $\mathbf{Z}^0$:

1) For $k = 1, \ldots, K$: $\theta_k^{(t+1)}|\mathbf{y}, \mathbf{Z}^{(t)}$ is sampled from $p(\theta_k|\mathbf{y}_k^{(t)})$.

2) For $i = 1, \ldots, n$: $Z_i^{(t+1)} | \mathbf{Z}_{-i}^{(t)}, y_i, \boldsymbol{\theta}^{(t+1)}$ is sampled from (2.22).

Note that K varies between the iterations. Basically only the θ_k of components of the infinite mixture (2.10), that are currently associated with at least one observation, are sampled. This sampler is applicable, if the integral in (2.22) can be solved analytically and if it is possible to sample from $p(\theta_k|\mathbf{y}_k)$, both being the case if G_0 is the conjugate distribution for $p(y|\theta)$. In that case it is also possible to analytically integrate over the θ_k to obtain the distribution

$$P(Z_i = k|\mathbf{Z}_{-i}, \mathbf{y}) \propto \begin{cases} \frac{n_{k,-i}}{\alpha+n-1} \int p(y_i|\theta) p(\theta|\mathbf{y}_{-i,k}) d\theta & , \quad k = 1, \ldots, K \\ \frac{\alpha}{\alpha+n-1} \int p(y_i|\theta) dG_0(\theta) & , \quad k = K+1 \end{cases} . \tag{2.23}$$

The Pólya urn Gibbs sampler reduces then to repeating step 2) with the distribution (2.23). Algorithms for non-conjugate G_0 are, for example, given by Neal (2000).

The second Gibbs sampler approximates the infinite mixture by a finite one. The random probability measure G in (2.17) is truncated at a large value N, so that $G = \sum_{k=1}^{N} \pi_k \delta_{\theta_k}$. The truncation is done by setting $V_N = 1$ in the stick-breaking construction (2.1) of the π_k and the resulting random probability measure is referred to as a truncated Dirichlet process (TDP). Ohlssen et al. (2007) choose N by considering $E(\sum_{k=N}^{\infty} \pi_k)$, the expected sum of the weights of the components that differ between the two processes. From (2.1) this is equal to $E(\prod_{k=1}^{N-1}(1 - V_k)) = (\frac{\alpha}{\alpha+1})^{N-1}$, since for the Dirichlet process $(1 - V_k)$ are independently $Beta(\alpha, 1)$ distributed. Ishwaran (2000) derives bounds on the total variation distance between the DP and TDP. The Gibbs sampler employing the TDP is referred to as blocked Gibbs sampler by Ishwaran and James (2001). It basically consists of the same steps as the Gibbs sampler for the finite mixture model in Subsection 2.1.4, except that

$\boldsymbol{\pi}|\mathbf{Z}$ is not sampled directly but via (2.12) and

$$V_k|\mathbf{Z} \sim Beta(1 + n_k, \alpha + \sum_{l=k+1}^{N} n_l) \ .$$

Recently, Papaspiliopoulos and Roberts (2008) proposed an MCMC sampler that avoids the need for truncating the stick-breaking construction of G. The idea is that sampling of π_{K^*} and θ_{K^*} for larger values of K^* is only started at that point where the number of components K first reaches K^* during the MCMC run.

2.2.4 Extensions of the Dirichlet Process

Its conjugacy and the availability of efficient computational methods have made the Dirichlet process a popular choice for modeling a single distribution with a Bayesian mixture. Several extensions of the DP, like hierarchical, nested and dependent DPs, have been proposed to model a collection of related distributions. As an example consider distributions of patient outcomes in several hospitals. For such a situation Teh et al. (2006) propose a hierarchical Dirichlet process where the distribution G_j of outcomes in the jth hospital follows a Dirichlet process

$$G_j|G_0 \sim DP(\alpha, G_0) \ ,$$

and the base measure G_0 itself also follows a DP

$$G_0 \sim DP(\gamma, H) \ .$$

This leads to the G_j being discrete random measures with the same atoms $\boldsymbol{\theta}$'s (drawn from H) and distinct but dependent $\boldsymbol{\pi}$'s. The hierarchical DP can be used to identify clusters of patients with similar outcomes over several hospitals, so that hospitals will share some of their clusters. In a related

approach, the nested Dirichlet process due to Rodriguez et al. (2008), the G_j are distributed as mixtures of Dirichlet processes via

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{G_k^*} \quad \text{and} \quad G_k^* \sim DP(\gamma, G_0) \ ,$$

where the π_{jk} follow the stick-breaking construction (2.12). In contrast to the hierarchical DP the nested DP assigns positive prior probability to G_j being equal to $G_{j'}$ for $j \neq j'$. It can thus be used for clustering hospitals with similar outcome distributions, while flexibly modeling the outcome distributions themselves.

The idea of replacing the atoms of an infinite mixture by stochastic processes is also useful for modeling related distributions G_x that are indexed by a a continuous covariate $x \in \mathcal{X}$. For example, Gelfand et al. (2005) develop a model for spatial data analysis with the outcome distribution at a location x being

$$G_x = \sum_{k=1}^{\infty} \pi_k \delta_{\Theta_k(x)}$$

where the Θ_k are 2-dimensional Gaussian processes. Alternatively Griffin and Steel (2006) allow the weights π_k to depend on x. Most important for these approaches is to define the random distributions so that they are continuous in the covariate, i.e., that $G_x \to G_{x_0}$ in distribution if $x \to x_0$. MacEachern (2000) gives a general theory for such dependent Dirichlet processes.

Chapter 3

Flexible Modeling with Bayesian Mixtures

3.1 Approximation of Distributions

When modeling the distribution of a random variable Y the assumption that P_Y belongs to a specific parametric family is often not justified. Results that show that Bayesian mixture models provide a suitable, more flexible alternative are given in this section. These results concern the possibility of approximating a distribution by mixtures and the accumulation of the posterior around the true distribution.

3.1.1 Density Estimation

From Figure 1.1 it becomes clear that mixtures of normals can approximate a wide variety of distributions. Escobar and West (1995) and Roeder and Wasserman (1997) therefore considered the use of Bayesian mixtures of normals for density estimation. An interesting question in this situation is whether it is possible to approximate any distribution on $\mathcal{Y}$ arbitrarily well

with a mixture of e.g., exponential family distributions.

To answer this question it has to be specified first, what one understands by "arbitrarily well". Let $f_0 \in \mathcal{M}$ be the true density of Y, with $\mathcal{M}$ being the space of all probability densities on $\mathcal{Y}$. To approximate f_0 arbitrarily well is then taken to mean that $\forall \varepsilon > 0$ there is a mixture with density $f_K = \sum_{k=1}^{K} \pi_k p(y|\theta_k)$, $K \leq \infty$, such that $d(f_0, f_K) < \varepsilon$, with $d(.,.)$ being a distance metric for densities. One common metric is the total variation metric

$$d_{TV}(f,g) = \int_{\mathcal{Y}} |f(y) - g(y)| dy \ .$$

Another often applied metric is the Prokhorov or weak metric $d_w(f,g)$. While this metric has a rather complicated form, (see, e.g., Gibbs and Su (2002)), it is mainly important because $d_w(f,g) \to 0$ is equivalent with the weak convergence of the underlying distributions P_f and P_g. As suggested by the names, the total variation metric is stronger than the weak metric as it can be shown that $d_w(f,g) \leq d_{TV}(f,g)$ for all $f, g \in \mathcal{M}$ (Huber and Ronchetti, 2009, p. 36).

In the context of approximating a prior distribution by a mixture of conjugate priors Diaconis and Ylvisaker (1985) show that any prior for an exponential family parameter can be approximated arbitrarily well in the weak sense. This implies that any distribution on $\mathbb{R}^p$ can be approximated arbitrarily well by a mixture of multivariate normals, any distribution on $(0,1)$ by a mixture of Betas and any distribution on $(0,\infty)$ by a mixture of Gammas. In their proof Diaconis and Ylvisaker (1985) show that any distribution can be approximated by a mixture of point masses and that any mixture of point masses can be approximated by a mixture of conjugate priors. This also demonstrates why an approximation in the weak sense is often not entirely satisfying, since it does not, for example, require that the approximation of

a continuous density is itself continuous.

Dalal and Hall (1983) consider the approximation of a prior density f_0 by a mixture of conjugate priors f_K in the stronger total variation sense. They show that f_0 can be approximated arbitrarily well by f_K if f_0 is bounded and continuous.

With these results it is clear that mixture models are generally flexible enough to be used for approximating distributions.

3.1.2 Consistency of Posterior Distribution

By placing a prior on either $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ or on a random mixing measure G in a Bayesian analysis of mixtures one also implicitly assigns a prior to the space of mixture densities $\mathcal{F} = \bigcup_{K=1}^{\infty} \mathcal{F}_K$, with $\mathcal{F}_K$ being the set of densities that are mixtures of K components. The space $\mathcal{F}$ is a subset of $\mathcal{M}$, the space of all probability densities on $\mathcal{Y}$. An additional question is then that of consistency, which means that given a sample $\mathbf{Y}_n = (Y_1, Y_2, \ldots, Y_n)'$ from a true distribution P_0 (with density f_0), does the posterior distribution on $\mathcal{F}$ accumulate around the true density $f_0 \in \mathcal{M}$?

Formally, let $A_\varepsilon = \{f : d(f_0, f) < \varepsilon\}$ be an ε-neighborhood of f_0 and let Π denote the induced probability distribution on $\mathcal{F}$. Consistency then means that

$$\forall \varepsilon > 0 : \quad \Pi(A_\varepsilon | \mathbf{Y}_n) \to 1 \quad a.s.\ P_0^\infty \ , \tag{3.1}$$

where P_0^∞ denotes repeated sampling from P_0. The neighborhood A_ε depends of course on the chosen metric d, e.g., there are weak and total variation neighborhoods. Depending on the type of neighborhoods for which (3.1) holds, it is spoken of weak or total variation consistency, the latter is also referred to as strong consistency.

An important role in proving consistency is played by Kullback-Leibler neighborhoods for which d is the Kullback-Leibler (KL) divergence

$$d_{KL}(f,g) = \int_{\mathcal{Y}} f(y) \log \frac{f(y)}{g(y)} dy \ . \tag{3.2}$$

It can be shown that $d_{KL}(f,g) \geq 0$, with equality only for $f \overset{a.e.}{=} g$, but that $d_{KL}(f,g) \neq d_{KL}(g,f)$ and that d_{KL} does not fulfill a triangle inequality. The Kullback-Leibler divergence is thus not a metric but has nevertheless proven to be useful in information theory, probability and statistics. Many of its properties are discussed in Cover and Thomas (1991). It holds that $d_{TV}(f,g)^2/2 \leq d_{KL}(f,g)$, so that Kullback-Leibler neighborhoods are smaller than weak and total variation ones.

A necessary condition for posterior consistency is that of support of the true density. This means that the prior puts mass in every neighborhood of f_0, formally: f_0 has support iff $\forall \varepsilon > 0 : \ \Pi(A_\varepsilon) > 0$. Different types of neighborhoods A_ε lead again to different types of support. Support basically means that f_0 can be approximated arbitrarily well and that some prior weight is assigned to the approximation.

Freedman (1963) shows that weak support does not necessarily imply weak consistency, whereas Schwartz (1965) proves that KL support is a sufficient condition for weak consistency. KL support is therefore an important property for which Wu and Ghosal (2008) provide an extensive treatment. For many kinds of mixture distributions they derive conditions on f_0 so that KL support holds. For mixtures of univariate normal distributions it is, for example, necessary for f_0 to be bounded and continuous and that $\int_{\mathbb{R}} |y|^{2+\delta} f_0(y) dy < \infty$ for some $\delta > 0$. Barron (1988) first established that total variation consistency follows from KL support and some additional regularity conditions on Π. As exemplified by Barron, Schervish, and Wasserman (1999) these conditions prevent the prior from assigning too much mass to

very rough densities. They present one set of conditions, others are provided by Ghosal, Ghosh, and Ramamoorthi (1999) and Walker (2004). Ghosal et al. (1999) also show that the regularity conditions are met by Dirichlet process mixtures of normals, thus establishing total variation consistency in this case. Roeder and Wasserman (1997) show that for a finite mixture of normal distributions total variation consistency holds if $f_0 \in \mathcal{F}$, i.e., if f_0 is itself a mixture of normals. If $f_0 \notin \mathcal{F}$, but in the KL support of $\mathcal{F}$, it is necessary to let K grow at a rate of $o(n/\log n)$ to obtain total variation consistency. These results show that Bayesian mixtures provide valid estimates of an unknown density f_0. And, as shown by Roeder and Wasserman (1997), they often outperform other commonly used methods, such as kernel density estimation, while being better interpretable.

3.1.3 Hierarchical Models

Frequently Bayesian mixtures are not used to model a distribution of observations directly but for distributions that are part of a larger statistical model. In commonly applied Bayesian hierarchical models, as described, for example, in Gelman et al. (2004), Bayesian mixtures can be used to model distributions on a higher level of the hierarchy. This is advantageous since parametric assumptions about distributions on these higher levels can be difficult to check, yet can have considerable influence on the obtained results. There are numerous applications using models of this kind, an early example being Bush and MacEachern (1996), who consider a randomized block design. They model the treatment effects parametrically, but use a Dirichlet process mixture for the distribution of the block effects. Another example is the work of Newton et al. (1996), who use a DPM to model the link function in a binary regression model. A recent review of related applications in

biostatistics is given by Dunson (2010). In the next section a detailed case study, concerning the flexible modeling of a random effects distribution in a logistic regression setting, is given.

3.2 Application: Goalkeepers' Performance in Saving Penalties

3.2.1 Background

In modern soccer, penalty shots are of vital importance. The world cup finals in 1990, 1994, and 2006, for example, were all decided by penalties. Special skills in saving penalties is commonly attributed to some goalkeepers. For example the German Wikipedia page on the penalty (http://de.wikipedia.org/wiki/Elfmeter, accessed 08.12.2009) asserts that there are some goalkeepers who are able to save more penalties than the average goalkeeper and gives a ranking of the German goalkeepers with the largest number of saved penalties. It is interesting from a statistical viewpoint that this ranking contains only the absolute number of saved penalties, not accounting for the number of potentially savable penalties for the respective goalkeeper.

Bornkamp, Fritsch, Kuß, and Ickstadt (2008) therefore approached the problem of ranking goalkeepers in a statistical more valid way. They considered all penalties from the German Bundesliga between August 1963 and May 2007. Here we repeat their analysis also including penalties from August 2007 to May 2009. In these 46 seasons a total number of 3907 penalties occurred. As we are focusing on the goalkeepers' ability to save penalties, we removed all penalties that missed the goal or hit goal-post or crossbar. This

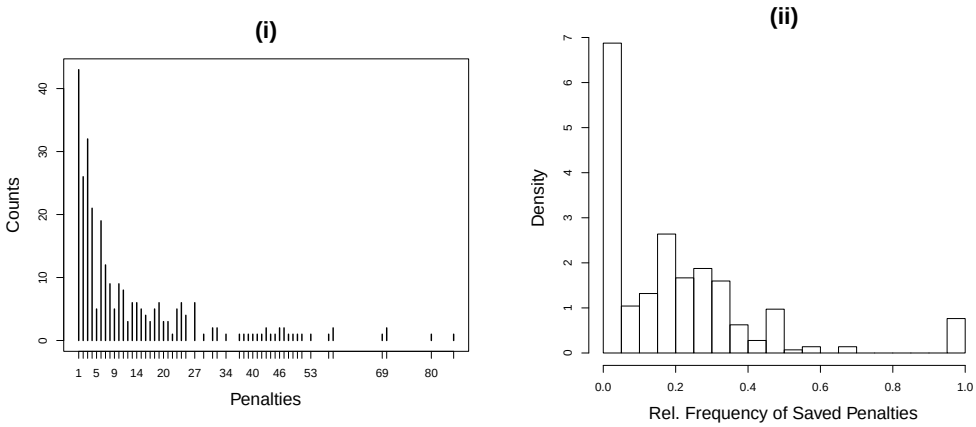

Figure 3.1: (i) Counts of penalties per goalkeeper and (ii) histogram of relative frequencies of saved penalties per goalkeeper.

resulted in 271 deletions with 3636 penalties remaining for the final analysis. Out of these 3636 penalties 736 were saved by the goalkeepers corresponding to a rate of 20.2%. In total 288 goalkeepers were involved in the 3636 penalties, many of them having been faced only with a small number of penalties (101 were involved in three or less penalties, see also Figure 3.1 (i)). Figure 3.1 (ii) shows the relative frequencies of saved penalties for all goalkeepers. The modes of the density at 0 and 1 are due to the goalkeepers that were involved in very few penalties and saved none or all. It is intuitively clear that a goalkeeper who was involved in only one single penalty during his career and saved this should not be considered the best penalty saver despite his 100% saving rate. Consequently, the relative frequency of saved penalties is a bad estimator of the "true" ability of the goalkeeper. A model also accounting for the number of penalties, that could potentially be saved, is needed.

3.2.2 Models

We will model the jth observed penalty of the ith goalkeeper as a realization of a Bernoulli random variable with probability ρ_{ij} that the goalkeeper saves

the penalty. This probability ρ_{ij} is modeled as a function of the ith goalkeeper and some additional covariates $\boldsymbol{x}_{ij}$. That is, we assume the model

$$\text{logit}(\rho_{ij}) = \gamma_i + \boldsymbol{\beta}'\boldsymbol{x}_{ij} \ , \quad i = 1, \ldots, 288 \ , \ j = 1, \ldots, n_i \ ,$$

where γ_i is the random effect of the ith goalkeeper, n_i is the number of penalties the ith goalkeeper was involved in, and $\boldsymbol{\beta}$ is the vector of regression coefficients for the covariates. Looking at Figure 3.1 (ii) it seems plausible that the distribution of goalkeeper effects is skewed or multimodal, even when the modes at 0 and 1 are ignored. The γ_i are therefore assumed to follow a mixture distribution. More specifically, a Dirichlet process mixture, as described in Subsection 2.2.2, is chosen. The components distributions are assumed to be normal with equal variance, so that the model is given by

$$\begin{aligned} \gamma_i|\mu_i, \sigma^2 &\sim N(\mu_i, \sigma^2) \\ \mu_i|G &\sim G \\ G &\sim DP(\alpha, N(0, 3.289)) \end{aligned} \tag{3.3}$$

The base measure $G_0 = N(0, 3.289)$ is chosen such that it is approximately uniform on the probability scale. The prior on the variance σ^2 of the normal components is chosen as a relatively vague $InvGa(1, 1/16)$ distribution and the distribution for the regression parameters $\boldsymbol{\beta}$ are chosen as independent vague uniform distributions. The parameter α of the Dirichlet process is set to $1/3$. As will be discussed in Subsection 5.2.2 this leads to a prior mean of $\approx$ 2.93 occupied clusters/components and virtually no prior mass on more than 8 components. This seems reasonable as we do not expect a large number of components for the penalty data.

To assess the merit of a flexible modeling of the random effects distribution via the proposed Dirichlet process model, we compare it to two more rigid

models via the deviance information criterion (DIC) (Spiegelhalter et al., 2002). The DIC is similar to AIC or BIC described in Subsection 2.1.6, but more suitable for hierarchical models. Defining $\boldsymbol{\rho}$ as the vector containing the probabilities ρ_{ij} the deviance is given by

$$D(\boldsymbol{\rho}|\mathbf{y}) = -2\sum_{i=1}^{288}\sum_{j=1}^{n_i} y_{ij}\log(\rho_{ij}) + (1-y_{ij})\log(1-\rho_{ij}) \ .$$

The DIC is then defined as $\overline{D(\boldsymbol{\rho}|\mathbf{y})}+p_D$, where $\overline{D(\boldsymbol{\rho}|\mathbf{y})}$ is the average deviance over the MCMC draws measuring the model fit and $p_D = \overline{D(\boldsymbol{\rho}|\mathbf{y})} - D(\bar{\boldsymbol{\rho}}|\mathbf{y})$ is an estimate of the "effective" number of parameters penalizing the model complexity ($\bar{\boldsymbol{\rho}}$ is the average of $\boldsymbol{\rho}$ over the MCMC iterations). For more details on the DIC we refer to Spiegelhalter et al. (2002).

The first model that will be used for comparison, is a model that does not allow for individual goalkeeper effects at all, leading to

$$\text{logit}(\rho_{ij}) = \mu_0 + \boldsymbol{\beta}'\boldsymbol{x}_{ij} \ ,$$

with a fixed common intercept μ_0. Hence, by comparing this model with the Dirichlet process model in terms of the DIC we will be able to quantify the improvement of modeling individual goalkeeper effects. The second model we use for a comparison is a parametric normal random effects model, which can be obtained by setting

$$\gamma_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

and using suitable vague hyper-priors for μ_0 and σ_0^2 (here we use $\mu_0 \sim \mathcal{N}(0, 3.289)$ and $\sigma_0^2 \sim InvGa(0.01, 0.01)$). By comparing the Dirichlet process model with this parametric model we will be able to quantify the improvement of a flexible modeling of the random effects distribution. Subsequently the two restricted models will be referred to as 'Intercept' and 'Normal', our proposed model will be termed the 'DP' model.

3.2.3 Choice of Covariates

The main aim of this analysis is to model the goalkeeper's effect on the probability of saving a penalty kick, but the effect of the scorer should also be taken into account. The logarithm of the number of taken penalties is chosen to represent the penalty takers' effect. For better interpretability the logarithm of base 2 is chosen. As home field advantage has an effect in many sports, the home field advantage of the goalkeeper is included as a binary covariate. To see whether there is a general time trend in the probability of saving a penalty, year is included as a covariate. "Year" here refers to a soccer season, which starts at the end of summer. A year effect could be due to improved techniques for saving or taking a penalty, e.g., Leininger and Ockenfels (2008) argue that it became more common to shoot a penalty into the middle of the goal after Neeskens succeeded with such a shot in the world cup final in 1974. This would lead to the penalty taker having more choices and thus would make it harder to save the penalty. The day of the season is also included as a covariate to account for possible time trends within a season.

3.2.4 Results

The models described above are fitted to the data using the OpenBUGS software version 3.0.3. Code implementing the blocked Gibbs sampler of Subsection 2.2.3 is adapted from Ohlssen et al. (2007). The truncation parameter N is set to 10. Larger values of N were tried as well, but did not influence results. For each model the MCMC sampler is run with two independent chains with a burn-in of 10,000 iterations followed by 100,000 iterations of which every 20th is kept. Trace plots of parameters did not indicate problems with

Model	$\overline{D(\rho\|y)}$	p_D	DIC
Intercept	3570.6	5.0	3575.6
Normal	3534.8	33.6	3568.4
DP	3530.4	37.3	3567.7

Table 3.1: Average deviance, effective number of parameters and DIC for the different models.

convergence of the chains and the results of the independent chains are similar. The results presented are based on the pooled draws of the independent chains, leading to a total number of 10,000 draws for each model.

First the overall fit of the models is compared with the DIC criterion. Table 3.1 shows the DIC and its components for the three models considered. Both the Normal and the DP model improve on the model with only an intercept, indicating some gain with the inclusion of a random effects distribution. The improvement is not very large, indicating that the probability of saving a penalty does not vary too much between goalkeepers. As it is more flexible, the DP model has a lower average deviance than the Normal model but also a larger number of effective parameters leading to a DIC that is only slightly lower.

Figure 3.2 shows the posterior expectation of the random effects distribution for the DP and Normal model. As might have been expected from the similar DICs, the distributions do not differ much between the two models. However, the more flexible DP model leads to a distribution with heavier tails than the one resulting from the Normal model. As can be seen from Figure 3.2 (ii) this distribution is not fully symmetric.

Next we take a look at the estimates for the goalkeepers' probabilities to

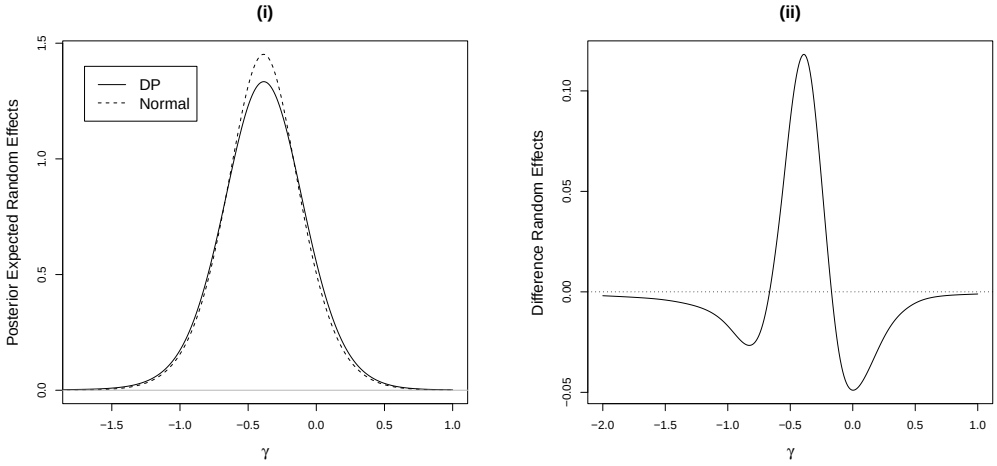

Figure 3.2: (i) Posterior expected random effects distribution γ for the Normal and DP model. (ii) Difference of Normal and DP posterior expected random effects distributions.

save a penalty that can be derived from the models. For this we consider

$$E\left(\frac{\exp(\gamma_i + \boldsymbol{\beta}'\boldsymbol{x}_{med})}{1+\exp(\gamma_i + \boldsymbol{\beta}'\boldsymbol{x}_{med})}\Big|\mathbf{y}\right),\ i = 1,\ldots,288, \tag{3.4}$$

the posterior expectation of the goalkeepers' probabilities to save a penalty kick when the covariates take their respective median values $\boldsymbol{x}_{med}$. The median values stand for a scorer with 9 taken penalties, the season 1985/86 and the 17th day of the season. The binary variable home field advantage is set to 0, representing no home field advantage for the goalkeeper. Figure 3.3 shows the posterior mean probabilities of the goalkeepers (from Equation (3.4)) for all goalkeepers smoothed by a kernel density estimate. Comparing Figure 3.3 (i) to the distribution of the relative frequencies in Figure 3.1 (ii) it can be seen that the probabilities are considerably shrunken towards each other. The range of estimates is only about 0.1. Figure 3.3 (ii) shows a close-up look at the distribution in (i), and as for the random effects distribution it can be seen that the estimates of the Normal and DP model

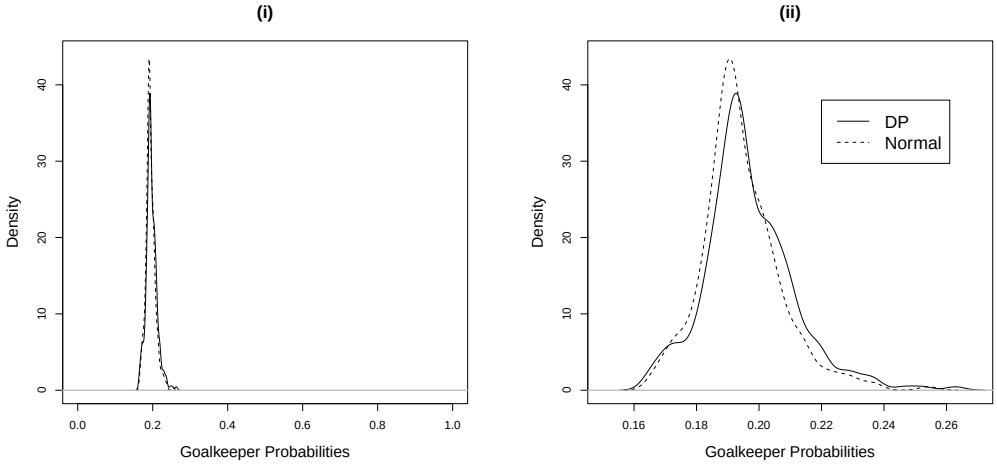

Figure 3.3: Posterior expected probabilities of saving a penalty for the Normal and DP model; (i) on the range [0,1] and (ii) on the range [0.15, 0.27].

differ mainly in the tails, with the DP model leading to more pronounced tails.

Regarding the question of identifying the best and worst keepers, the tails of the distribution are of importance. As the DP model is more flexible in the tails it is used to determine a ranking of the keepers. In performing the ranking, part of which is given in Table B.1 in the appendix, we rely on the recommendations of Lin et al. (2006) who argue that ranking should be based on the posterior expectation of the rank rather than the posterior expected effect. This explains the fact that in some cases a goalkeeper with a higher rank nevertheless has a higher posterior expected probability of saving a penalty. An example of this is given by Robert Enke and Jean-Marie Pfaff having rank 2 and 3 and estimated saving probabilities of 24.3% and 24.9%, respectively. The top goalkeeper with an estimated saving probability of 26.0% is Rudolf Kargus. With 23 saved penalties out of 70 he is also the goalkeeper with the highest absolute number of saved penalties.

Several other interesting observations arise from the ranking in Table

B.1. Goalkeepers' estimated saving probabilities are not really different, with the best keeper having 26.0% and the worst keeper having 16.2%, yielding only a 10%-points difference. Moreover, the credible intervals for the saving probabilities are seen to be pretty large, credible intervals for the best and the worst keeper overlap considerably. As such, saving capabilities are rather similar across goalkeepers. It is nevertheless surprising, that two German goalkeepers who are thought to be penalty specialists (Oliver Kahn and Jens Lehmann) rank relatively low (and right next to each other), indicating that both of them perform rather badly in penalty saving. This is probably due to the perception of the German expertise in penalty shoot-outs in recent tournaments, with Kahn and Lehmann playing prominent roles on these occasions.

The hierarchical model leads to a considerable shrinkage effect. This is demonstrated by Michael Melka and Gerhard Teupel as two representatives of the goalkeepers who were faced with only one single penalty during their career in the German Bundesliga. Michael Melka, who saved this single penalty (thus having an observed 100% saving rate), has an estimated saving probability of only 20.4%. Gerhard Teupel, not saving this single penalty (resulting in an observed 0% saving rate) estimated saving probability is 19.2%, not very different from Melka's probability. It is somewhat surprising to see a well-known goalkeeper like Sepp Maier rank so low. This is a direct consequence of the shrinkage effect of the random effects model: As can be seen in Table B.1, only goalkeepers who were involved in many penalties can rank at the top or the bottom of the list, while the goalkeepers with fewer penalties are all in the middle of the ranking. This is reasonable from a statistical point of view, as we can only make statistically accurate estimates for keepers with many penalties, while those with few penalties are shrunken towards the overall

Covariate	OR with 95% CI
Scorer	0.752 [0.710, 0.795]
Home Field Advantage	0.943 [0.784, 1.131]
Year	0.805 [0.571, 1.101]
Day of Season	0.930 [0.706, 1.203]

Table 3.2: Estimated odds ratios with 95% credible intervals in the DP model. For the penalty taker odds ratio is for a scorer with twice the number of penalties. The odds ratio for year compares the last to the first year, which is also the case for day of the season.

mean. This shrinkage effect should be kept in mind, when interpreting the ranking of goalkeepers from an application viewpoint.

Finally, we consider the effects of the covariates. Since a logistic regression model is fitted, $\exp(\beta_j)$ can be interpreted as the change in the odds of the event, if the jth covariate is risen by 1. Table 3.2 shows the estimated odds ratios for the DP model. As the credible interval for the odds ratio of the scorer effect does not contain 1 there is strong evidence that a scorer that has taken more penalties reduces the goalkeeper's probability of saving the penalty. This is a reasonable result, since players that are known to be good penalty takers are probably chosen more often to take a penalty kick. As the scorer effect is given on the log2 scale, the odds ratio can be interpreted as follows: Faced with a scorer that scored twice as many penalties, the goalkeeper's odds of saving is multiplied by 0.752. For all the other covariates, 1 is inside the credible interval. This implies that there is no evidence for a home field advantage for the goalkeeper. Additionally, evidence can neither be found for an overall time trend or a time trend within seasons. These conclusions are also obtained for the other two models.

Chapter 4

Introduction to Cluster Analysis

> *Gallia est omnis divisa in partes tres, [...]. Hi omnes lingua, institutis, legibus inter se differunt.*
>
> *All Gaul is divided into three parts, [...]. All these differ from each other in language, customs and law.*
>
> Julius Caesar, Commentarii de Bello Gallico

Cluster analysis is the attempt to group previously unstructured data so that the observations in a group are more similar to each other than to observations from other groups. As the above quote shows people have informally defined such groupings for a long time. And they continue to do so, as the conscious or unconscious grouping of objects is an important part of learning and structuring new knowledge. Usually there is not one single way in which a given set of objects should be grouped or clustered. A useful analogy is the sorting of books in a library: there is more than one meaningful way to do it. The way a set of objects should be grouped depends on the aspects of the

objects that are of interest and how one defines similarity between objects. Statistics can help to formalize and automate such a process, either by defining criteria that should be optimized by a clustering or by formulating an underlying probability model. The latter is referred to as model-based cluster analysis. In the next chapter the use of Bayesian mixtures for model-based cluster analysis will be discussed in detail.

Cluster analysis started to be developed and applied in evolutionary biology and psychology and is nowadays used in a wide variety of fields, ranging from astronomy (classification of stars), geography (grouping of regions), chemistry (grouping of compounds) to marketing (market segmentation). Overviews are provided by Kaufman and Rousseeuw (1990), Everitt et al. (2001) and Mirkin (2005). A recent surge of interest in cluster methods is due to applications in bioinformatics, after a seminal paper by Eisen et al. (1998) demonstrated that the clustering of microarray gene expression measurements can identify groups of functionally related genes. See, for example, Podwojski et al. (2009) for a recent bioinformatics application. In the last years many applications also came up in the context of data mining (Mirkin, 2005), e.g., the clustering of text documents in large databases. In the next section a short introduction to some of the classical methods of cluster analysis will be given and it will be shown that many common clustering criteria can be better understood in the context of mixture models. In Section 4.2 distance measures between clusterings will be introduced as these will be important for our study of cluster analysis with Bayesian mixtures.

4.1 Classical Methods

Formally a clustering $\mathcal{C}$ is a partition of a set S of n objects or observations into K subsets $C_1, \ldots, C_K$ such that $C_k \cap C_{k'} = \emptyset$ for $k \neq k'$ and $\bigcup_{k=1}^{K} C_k = S$. Cluster methods can be roughly grouped into partitioning methods and hierarchical methods (this is, by the way, a clustering of clustering methods). The former try to find an optimal clustering with regard to some criterion for a fixed number of groups K, while the latter form a sequence of clusterings $\mathcal{C}^{(K)}$ for $K = 1, \ldots, n$, such that $\mathcal{C}^{(n)}$ has each observation in a singleton cluster and $\mathcal{C}^{(K-1)}$ is obtained from $\mathcal{C}^{(K)}$ by merging two of its clusters.

4.1.1 Partitioning Clustering

An allocation vector $\mathbf{Z}$ as employed in Chapter 2 uniquely defines a clustering $\mathcal{C}(\mathbf{Z})$ (the opposite is not true, see Chapter 2.1.3 and Chapter 5.3). For ease of notation in this subsection a clustering will be referred to by an allocation vector $\mathbf{Z}$. The goal is then to find a clustering $\mathbf{Z}$ with K groups that optimizes a criterion $c(\mathbf{Z})$. Common criteria for observations $y_1, \ldots, y_n$ from $\mathbb{R}^p$ are based on an ANOVA like decomposition of the total variability $T = \sum_i (y_i - \bar{y})(y_i - \bar{y})'$ of the data into

$$
\begin{aligned}
T &= W(\mathbf{Z}) + B(\mathbf{Z}) \qquad \text{where} \\
W(\mathbf{Z}) &= \sum_{i=1}^{n} (y_i - \bar{y}_{Z_i})(y_i - \bar{y}_{Z_i})' \\
&= \sum_{k=1}^{K} \sum_{i:Z_i=k} (y_i - \bar{y}_k)(y_i - \bar{y}_k)' \quad \text{and} \\
B(\mathbf{Z}) &= \sum_{i=1}^{n} (\bar{y}_{Z_i} - \bar{y})(\bar{y}_{Z_i} - \bar{y})' \\
&= \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})(\bar{y}_k - \bar{y})' \ .
\end{aligned}
$$

It is then desirable to have the within-cluster variability $W(\mathbf{Z})$ as small as possible (and accordingly the between-cluster variability $B(\mathbf{Z})$ as large as possible). The K-means method proposed by MacQueen (1967) attempts to minimize the squared Euclidean distance of each observation to its cluster mean, which is the same as minimizing $\mathrm{tr}(W(\mathbf{Z}))$. Friedman and Rubin (1967) considered the minimization of the determinant $|W(\mathbf{Z})|$.

Scott and Symons (1971) (see also Frühwirth-Schnatter (2006, pp. 207-210)) showed that these heuristic criteria arise from maximizing $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta})$ given in (2.4) with respect to $\mathbf{Z}$ and $\boldsymbol{\theta}$ for a mixture of normal distributions. Note that this is not the same as maximizing the finite mixture likelihood (2.3) and is referred to as classification likelihood. For a mixture of normals

$$\log p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto -\frac{1}{2}\sum_{k=1}^{K}\left(n_k \log|\Sigma_k| + \sum_{i:Z_i=k}(y_i-\mu_k)'\Sigma_k^{-1}(y_i-\mu_k)\right) .$$

It is a well known result in multivariate statistics that maximizing $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with fixed $\mathbf{Z}$ reduces to maximizing $p(\mathbf{y}|\mathbf{Z}, \widehat{\boldsymbol{\mu}}(\mathbf{Z}), \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$, where $\widehat{\mu}_k = \bar{y}_k$. $p(\mathbf{y}|\mathbf{Z}, \widehat{\boldsymbol{\mu}}(\mathbf{Z}), \boldsymbol{\Sigma})$ is then proportional to

$$-\frac{1}{2}\left(\sum_{k=1}^{K} n_k \log|\Sigma_k| + \mathrm{tr}(W_k(\mathbf{Z})\Sigma_k^{-1})\right) , \tag{4.1}$$

with $W_k(\mathbf{Z}) = \sum_{i:Z_i=k}(y_i-\bar{y}_k)(y_i-\bar{y}_k)'$ being the variability in cluster k. For a mixture with spherical components of equal variance $\Sigma_k = \sigma^2 I$ and (4.1) reduces to

$$-\frac{1}{2}\left(np\log\sigma^2 + \frac{1}{\sigma^2}\mathrm{tr}(W(\mathbf{Z}))\right) ,$$

which for fixed $\mathbf{Z}$ is maximized by $\widehat{\sigma}^2 = \mathrm{tr}(W(\mathbf{Z}))/n$. To find an optimal clustering one then has to maximize

$$c(\mathbf{Z}) = -\frac{1}{2}np\log\mathrm{tr}(W(\mathbf{Z})) , \tag{4.2}$$

which is the same as minimizing $\mathrm{tr}(W(\mathbf{Z}))$. It can be shown similarly that minimizing $|W(\mathbf{Z})|$ arises from assuming a homoscedastic mixture $\Sigma_k = \Sigma$. The approach can of course be extended to other assumptions about the variance structure of the mixture. A spherical mixture with different variances $\Sigma_k = \sigma_k^2 I$ leads to the minimization of

$$\prod_{k=1}^{K} \mathrm{tr}\left(\frac{W_k(\mathbf{Z})}{n_k}\right)^{n_k}$$

with respect to $\mathbf{Z}$. For heteroscedastic mixtures with unconstrained covariance matrices Σ_k the expression to be minimized is

$$\prod_{k=1}^{K} \left|\frac{W_k(\mathbf{Z})}{n_k}\right|^{n_k} \quad .$$

The above criteria have been derived by maximizing $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that it is implicitly assumed that $p(\mathbf{Z}|\boldsymbol{\pi})$ is fixed. It can be seen from (2.5) that this is equivalent to assuming that $\boldsymbol{\pi} = (1/K, \ldots, 1/K)'$. Symons (1981) considered maximizing $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})p(\mathbf{Z}|\boldsymbol{\pi})$. For a given $\mathbf{Z}$ this is maximized by $\widehat{\pi}_k = n_k/n$, without affecting $\widehat{\mu}_k$ and $\widehat{\Sigma}_k$. Including maximization with respect to $\boldsymbol{\pi}$ then changes (4.2) to

$$c(\mathbf{Z}) = -\frac{1}{2} np \log \mathrm{tr}(W(\mathbf{Z})) + \sum_{k=1}^{K} n_k \log n_k \quad .$$

Similar criteria can be obtained for the other assumptions about the variance. These results give an explanation for the empirical finding that the K-means algorithm tends to produce spherical clusters of similar size.

4.1.2 Hierarchical Clustering

Hierarchical clustering produces a sequence of nested clusterings for $K = 1, \ldots, n$. One distinguishes between agglomerative and divisive approaches

depending on whether the method starts with n clusters which are successively merged or with one cluster, that is then successively split. Agglomerative methods are the more commonly applied. The clusters are constructed based on the matrix of pairwise distances of the observations $d(i, j)$, $i, j = 1, \ldots, n$. d is usually a metric, but might also be a dissimilarity measure that does not fulfill all requirements of a metric. All agglomerative methods start by merging the two objects i' and j' that have minimal distance $d(i', j')$ into a cluster. The distances between objects $d(i, j)$ are then used to construct distances between clusters $d(C_k, C_{k'})$, with the methods differing in the way this is done. For the average linkage method by Sokal and Michener (1958) the distance between cluster k and k' is

$$d(C_k, C_{k'}) = \frac{1}{n_k n_{k'}} \sum_{i \in C_k, j \in C_{k'}} d(i, j) \ , \tag{4.3}$$

where n_k denotes the number of objects in cluster k. The distance of the two clusters is thus the average distance of an object in cluster k to an object in cluster k'. After merging two clusters the distances of the obtained new cluster to the other clusters are computed. It is not necessary to resort to the object-wise distances, as the cluster-wise distances can be updated. For average linkage the updating formula is given by

$$d(C_k \cup C_{k'}, C_{k''}) = \frac{n_k}{n_k + n_{k'}} d(C_k, C_{k''}) + \frac{n_{k'}}{n_k + n_{k'}} d(C_{k'}, C_{k''}) \ . \tag{4.4}$$

The two closest clusters with respect to the updated distances are merged in the next step until all observations are in one cluster. Other common agglomerative methods are single linkage and complete linkage, where the distance of two clusters is the minimum respectively maximum pairwise distance between objects in the two clusters. The method of Ward (1963) is the hierarchical equivalent to the K-means method. It is used for Euclidean

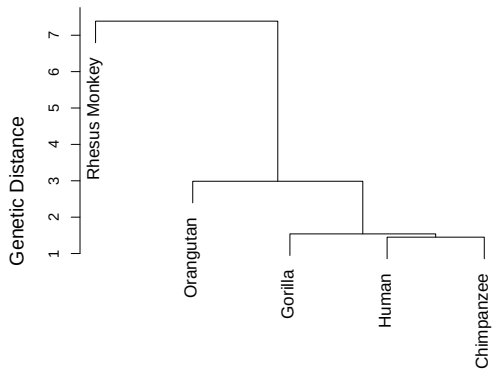

Figure 4.1: Dendrogram of genetic distances.

distances $d(i, j)$ and at each step merges two clusters such that the increase in $\text{tr}(W(\mathbf{Z}))$ is minimal ($\text{tr}(W(\mathbf{Z}))$ cannot decrease when merging clusters). Lance and Williams (1966) give a general updating formula similar to (4.4) that applies to single and complete linkage. Wishart (1969) showed that there is also an updating formula for Ward's method.

A hierarchical clustering is usually visualized by a dendrogram, see Figure 4.1 for an illustrative example. Average linkage is applied to genetic distances between Human, the three great apes and the Rhesus monkey (data taken from Mirkin (2005, p. 5)). The dendrogram visualizes the distance at which mergers have taken place. The distance between two objects is given by the vertical distance at which their branches coincide. These distances possess the *ultrametric* property $d(i, j) \leq \max\{d(i, l), d(j, l)\}$, which is a stronger version of the triangle inequality. The property implies that two of the distances $d(i, j)$, $d(i, l)$ and $d(j, l)$ are equal and that the third is not larger than the other two. In the genetic example shown in Figure 4.1 the whole hierarchy is of interest as it represents an evolutionary tree. In other situations hierarchical

clustering is used as an informal method to select the number of clusters K, e.g., by cutting the dendrogram at the point where the vertical distance to the next merger is maximal. For the genetic data this would imply to form one cluster containing Human and the great apes and one containing only the Rhesus monkey.

4.2 Similarity Measures for Clusterings

It is often of interest to have a measure of how close two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of the same objects are. This might be the case because one wants to compare the results of several clustering algorithms or because a clustering is to be compared to a known partition, e.g., in a validation study of clustering methods. Here we will consider three classes of similarity measures: measures based on pairs of observations, based on cluster matching and based on information theory.

The first class of similarity measures is concerned with the treatment of pairs of observations in the two clusterings. Each of the $\binom{n}{2}$ pairs of objects belongs to one of the four categories:

(i) objects are in the same cluster in $\mathcal{C}$ and in the same cluster in $\mathcal{C}'$,

(ii) objects are not in the same cluster in $\mathcal{C}$ and not in the same cluster in $\mathcal{C}'$,

(iii) objects are in the same cluster in $\mathcal{C}$ and not in the same cluster in $\mathcal{C}'$,

(iv) objects are not in the same cluster in $\mathcal{C}$ and in the same cluster in $\mathcal{C}'$.

Let $n(\text{x})$ denote the number of pairs of type (x), then $n(\text{i}) + n(\text{ii}) = A$ is sometimes called the number of agreements. Vice versa $n(\text{iii}) + n(\text{iv}) = D$

is called the number of disagreements. Note that $n(\text{i})+n(\text{ii})+n(\text{iii})+n(\text{iv})= A + D = \binom{n}{2}$. Rand (1971) proposed $R(\mathcal{C},\mathcal{C}') = A/\binom{n}{2}$, the proportion of agreements, as a similarity measure for clusterings. The Jaccard index $J(\mathcal{C},\mathcal{C}') = n(\text{i})/[n(\text{i}) + n(\text{iii}) + n(\text{iv})]$ is similar but does not count pairs of type (ii). Because of the fixed number of pairs, they are, however, still implicitly included. The index of Fowlkes and Mallows (1983) does also not explicitly consider the pairs of type (ii) and is given by

$$FM(\mathcal{C},\mathcal{C}') = \frac{n(\text{i})}{\sqrt{[n(\text{i}) + n(\text{iii})][n(\text{i}) + n(\text{iv})]}} \ .$$

All these indices take values in [0,1] and are equal to 1 if the two clusterings are identical. $1-R(\mathcal{C},\mathcal{C}')$ can be shown to be a metric for clusterings (Mirkin, 1996). The Rand index can be interpreted as the probability that a randomly chosen pair of objects is treated the same in both clusterings. A probabilistic interpretation is also available for the Fowlkes and Mallows index: it is the geometric mean of the probability that a pair of observations is in one cluster in $\mathcal{C}$ given it is in one cluster in $\mathcal{C}'$ and the same probability with the role of $\mathcal{C}$ and $\mathcal{C}'$ interchanged.

Hubert and Arabie (1985) recognized that for the Rand index the number of expected chance agreements between the two clusterings depends heavily on the number of groups in each clustering, their sizes, and the overall number of observations. To overcome this problem they considered the contingency table of two clusterings shown in Table 4.1 and proposed an adjusted Rand index, where the index is corrected for its expected value under the assumption of random sampling of the $n_{kk'}$ from fixed marginal sizes $n_{k\cdot}$ and $n_{\cdot k'}$, i.e., assuming a generalized hypergeometric distribution for the contingency table. The adjusted Rand has the usual form of an index corrected for chance:

$$\frac{\text{Index - Expected Index}}{\text{Maximum Index - Expected Index}} \ .$$

		Clustering $\mathcal{C}'$			
	Cluster	C'_1	$\cdots$	$C'_{K'}$	$\sum$
	C_1	n_{11}	$\cdots$	$n_{1K'}$	$n_{1.}$
Clustering $\mathcal{C}$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	C_K	n_{K1}	$\cdots$	$n_{KK'}$	$n_{K.}$
	$\sum$	$n_{.1}$	$\cdots$	$n_{.K'}$	n

Table 4.1: The contingency table of two clusterings.

It has a maximum value of 1 and its value is 0 if the index equals its expected value. Negative values are possible, but uninteresting as they indicate less agreement than expected by chance. Hubert and Arabie (1985) derive the following formula for the adjusted Rand index:

$$AR(\mathcal{C},\mathcal{C}') = \frac{\sum_{k,l}\binom{n_{kk'}}{2} - \sum_k \binom{n_{k.}}{2}\sum_{k'}\binom{n_{.k'}}{2}/\binom{n}{2}}{\frac{1}{2}[\sum_k \binom{n_{k.}}{2} + \sum_{k'}\binom{n_{.k'}}{2}] - \sum_k \binom{n_{k.}}{2}\sum_{k'}\binom{n_{.k'}}{2}/\binom{n}{2}} \quad . \tag{4.5}$$

Milligan and Cooper (1986) compared the four above mentioned indices in a simulation study. They fitted hierarchical clusterings to data without cluster structure (i.e., drawn from uniform distributions over a region) and computed the similarity of the clusterings across the hierarchy with a fixed reference clustering. Since there is no relation to the reference clustering the indices should on average be constant over the hierarchy. Milligan and Cooper (1986) found this only to be the case for the adjusted Rand index, whereas the Rand index tended to increase and the Jaccard and Fowlkes and Mallows indices tended to decrease with the number of clusters in the hierarchical clustering. They also found that especially the Rand index has a very large variance, which is not the case for the adjusted Rand. Since they could also show that the adjusted Rand is able to recover a cluster structure present in the data they recommend it as measure for the comparison of clusterings.

Another possibility for measuring similarity of clusterings is by matching the clusters in both clusterings and computing classification rates. Meilă and Heckerman (2001) give such a matching measure by

$$M(\mathcal{C}, \mathcal{C}') = \frac{1}{n} \max_{\rho} \sum_{k=1}^{K} n_{k\rho(k)} \ ,$$

where, without loss of generality, $K \leq K'$ and ρ is an injective mapping from $\{1, \ldots, K\}$ to $\{1, \ldots, K'\}$. So the clusters in both clusterings are first (partially) matched and the percentage of correctly classified observations is computed. The similarity measure is then obtained by maximizing the classification rate over all matchings. One problem with a matching similarity is that it ignores what happens to the unmatched part of each cluster. As pointed out by Meilă (2007) quite different clusterings $\mathcal{C}'$ result, for example, if the unmatched part of each cluster in $\mathcal{C}$ is spread equally over the clusters in $\mathcal{C}'$ or if it is assigned to only one other cluster. These clusterings will have the same matching similarity to $\mathcal{C}$ although intuitively the latter should be more similar. A simulation study by Steinley (2004) showed that the matching similarity decreases only slowly when the overlap between two clusterings is decreased and only rarely attains values less than 0.3. He also finds that the adjusted Rand has a superior performance.

Meilă (2007) derives a (dis-)similarity measure based on information theory. Each clustering $\mathcal{C}$ defines a discrete random variable C with $P(C = k) = n_k/n$. The "variation of information"-distance of Meilă (2007) between clusterings $\mathcal{C}$ and $\mathcal{C}'$ is then given by sum of the conditional entropies of the associated random variables C and C', i.e.,

$$VI(\mathcal{C}, \mathcal{C}') = H(C|C') + H(C'|C) \ ,$$

where the conditional entropy is given by

$$H(C|C') = -\sum_{k'=1}^{K'} P(C' = k') \sum_{k=1}^{K} P(C = k|C' = k') \log P(C = k|C' = k') \ .$$

It is a measure of the uncertainty in C that is not explained by C', see Cover and Thomas (1991, Chapter 2) for details. Meilă (2007) shows that the VI-distance has many desirable properties. First, it is a metric for the space of clusterings, i.e., $VI(\mathcal{C}, \mathcal{C}') \geq 0$ with equality if and only if $\mathcal{C} = \mathcal{C}'$, $VI(\mathcal{C}, \mathcal{C}') = VI(\mathcal{C}', \mathcal{C})$ and $VI(\mathcal{C}, \mathcal{C}') \leq VI(\mathcal{C}, \mathcal{C}'') + VI(\mathcal{C}', \mathcal{C}'')$. In addition it is n-invariant, meaning that the value of VI only depends on the relative sizes of the clusters and not directly on the number of observations n. If, for example, all entries in Table 4.1 are multiplied by a constant the criterion is not changed. The VI-distance also has the property of *convex additivity*, which means that if $\mathcal{C}'$ and $\mathcal{C}''$ are each obtained from $\mathcal{C}$ by further splitting the clusters of $\mathcal{C}$, the VI-distance of $\mathcal{C}'$ and $\mathcal{C}''$ is given by

$$VI(\mathcal{C}', \mathcal{C}'') = \sum_{k=1}^{K} P(C = k) VI(\mathcal{C}'_k, \mathcal{C}''_k) \ ,$$

where $\mathcal{C}'_k$ and $\mathcal{C}''_k$ are the partitions of C_k. Convex additivity implies that if cluster C_k is split to obtain a new clustering $\mathcal{C}'$ the distance to the original clustering does only depend on the size of C_k and the way it is split and not on the way the rest of the data is clustered. Meilă (2007) discusses in detail which of the mentioned properties are fulfilled by the other similarity measures, the adjusted Rand index, for example, does not possess convex additivity.

Sometimes the mutual information $MI(C, C')$, which is related to the VI-distance by

$$VI(\mathcal{C}, \mathcal{C}') = H(C) + H(C') - 2 \cdot MI(C, C') \ , \tag{4.6}$$

is used as a similarity measure for clusterings. H in (4.6) denotes the (unconditional) entropy given by $H(C) = -\sum_{k=1}^{K} P(C = k) \log P(C = k)$. Since the mutual information is not a metric for clusterings the use of the VI-distance should generally be preferred.

Treppmann (2010) repeated the simulation study of Milligan and Cooper (1986) mentioned above. In addition to the similarity measures based on pairs she included the matching measure and the VI-distance. Both new measures successfully identify cluster structure when it is present. When applied to the hierarchical clustering of noise data, however, both criteria fail to be on average constant across the hierarchy. This is especially problematic for the VI-distance. Treppmann (2010) found that the adjusted Rand index still gives the best results.

In the following we will employ both the adjusted Rand index and the VI-distance for the comparison of clusterings. The former is used because of its good performance in simulation studies and the latter because of its many reasonable theoretical properties.

Chapter 5

Cluster Analysis with Bayesian Mixtures

Due to their computational efficiency the classical methods of cluster analysis described in Section 4.1 remain popular, although it is difficult to assess the statistical properties of the solutions provided by these methods. It is, for example, hard to quantify the uncertainty of the allocation of an observation to a specific group or the probability that two observations belong to the same group. These quantities can easily be estimated using model-based cluster methods based on mixtures. Among the advantages of a Bayesian approach to these models are the ones discussed in Subsection 2.1.5 and that MCMC algorithms allow to take the uncertainty in parameters better into account than single estimates. Bayesian mixtures can also be extended to fairly complex models. As discussed in Chapter 2, one can, for example, estimate the number of clusters K at the same time as the other parameters by either the reversible jump algorithm (Richardson and Green, 1997) or by infinite mixture models.

Other implementations of complex Bayesian cluster models allow for out-

lier detection (Quintana and Iglesias, 2003), simultaneous clustering and variable selection (Kim et al., 2006; Tadesse et al., 2005), improving the power of multiple testing by clustering correlated observations (Dahl and Newton, 2007), or clustering transcription factor binding motifs of varying width (Jensen and Liu, 2008).

In Section 5.1 it will be discussed under which conditions a Bayesian mixture can be used for cluster analysis and in Section 5.2 priors that Bayesian mixtures induce on important quantities, like the number of clusters and the clusterings itself, will be given. In Section 5.3 and 5.4 existing approaches for clustering inference with fixed and varying number of groups K are discussed and new methods are developed for both situations. Section 5.5 gives applications of cluster analysis with Bayesian mixtures to simulated and real data.

5.1 Conditions for Cluster Analysis: An Example

It has been seen in Chapter 3 that a mixture model can be used as an approximation to (almost) any desired distribution. In this section it will be considered under what conditions a mixture model can be used for cluster analysis. Assume that $Y_1, \ldots, Y_n$ are independent random variables with a finite mixture distribution of the form (2.1). For now we will assume that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are known. Then cluster analysis requires to make inferences about the $Z_i|Y_i$, with distributions given by (2.2). To obtain a useful clustering it should be possible to assign most Z_i with high probability to one of the components 1 to K, i.e., $\max_j P(Z_i = j|Y_i = y_i)$ should be large for most i. To infer a group structure it is also desirable if the Z_i that can be

assigned with high probability are not all assigned to the same component.

To formalize these properties of a mixture model we define Z_i to be *uncertain* if $\max_j P(Z_i = j|Y) < u$. To be useful for clustering a mixture model should fulfill the conditions:

1. $q = P_Y(Z \text{ uncertain})$ should be small.

2. $P_Y(Z = k|Z \text{ not uncertain}) > 0$ for more than one k.

It holds that

$$\begin{aligned} q &= P_Y(\max_j P(Z = j|Y) < u) \\ &= P_Y\left(\frac{\max_j \pi_j p(Y|\theta_j)}{\sum_{k=1}^K \pi_k p(Y|\theta_k)} < u\right) \\ &= \int_U p(y|\boldsymbol{\theta}, \boldsymbol{\pi})dy \qquad \text{with} \quad U = \left\{ y \left| \frac{\max_j \pi_j p(y|\theta_j)}{\sum_{k=1}^K \pi_k p(y|\theta_k)} < u \right. \right\} \ . \end{aligned}$$

As a simple example consider a univariate location mixture of two normals with $\sigma^2 = 1$, i.e.,

$$p(y|\boldsymbol{\mu}, \pi) = \pi\phi(y|\mu_1, 1) + (1 - \pi)\phi(y|\mu_2, 1) \ .$$

The region U where an observation is likely to be of either component and thus uncertain is in this case an interval $[l_1, l_2]$, that will usually lie inside $[\mu_1, \mu_2]$ (assuming w.l.o.g. that $\mu_1 < \mu_2$). Note, however, that for π close to 0 l_1 can be smaller than μ_1 and for π close to 1 l_2 can be larger than μ_2. The probability of an uncertain Z is then given by $q = \int_{l_1}^{l_2} p(y|\boldsymbol{\mu}, \pi)dy$. The lower limit l_1 can be found by equating

$$\frac{\pi\phi(l_1|\mu_1, 1)}{\pi\phi(l_1|\mu_1, 1) + (1 - \pi)\phi(l_1|\mu_2, 1)} = u \ ,$$

which is solved by

$$l_1 = \bar{\mu} - \frac{\log\left(\frac{u}{1-u}\frac{1-\pi}{\pi}\right)}{\mu_2 - \mu_1} \ ,$$

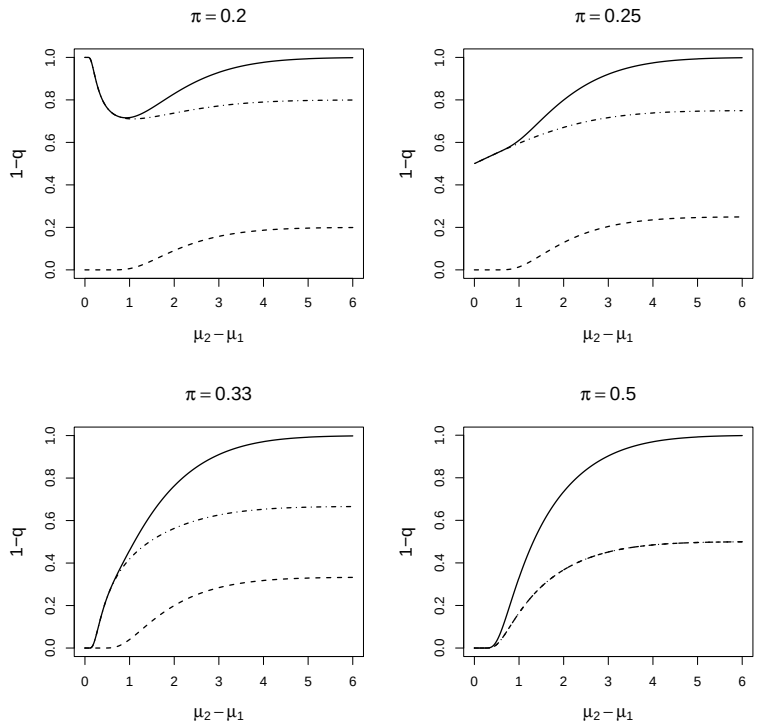

Figure 5.1: Probability that cluster membership Z is not uncertain. Solid line: Total probability $1-q$; Dashed (dash-doted) line: part of probability due to cluster 1 (2). Uncertainty threshold $u = 0.75$.

with $\bar{\mu} = (\mu_1 + \mu_2)/2$. Similarly the upper limit l_2 is given by

$$l_2 = \bar{\mu} - \frac{\log\left(\frac{1-u}{u}\frac{1-\pi}{\pi}\right)}{\mu_2 - \mu_1} \quad . \tag{5.1}$$

The probability of an uncertain Z is then

$$\begin{aligned} q &= \int_{l_1}^{l_2} \pi\phi(y|\mu_1, 1) + (1-\pi)\phi(y|\mu_2, 1)dy \\ &= \pi\left[\Phi(l_2 - \mu_1) - \Phi(l_1 - \mu_1)\right] + (1-\pi)\left[\Phi(l_2 - \mu_2) - \Phi(l_1 - \mu_2)\right] \quad . \end{aligned}$$

Figure 5.1 shows $1-q$ as a function of $\mu_2 - \mu_1$ for different values of π, where the uncertainty threshold u is taken to be 0.75. For $\pi = 0.2$ the probability of Z not being uncertain is close to 1 for small values of $\mu_2 - \mu_1$ and first

decreases when $\mu_2 - \mu_1$ gets larger until it rises again. A reason for this can be seen from (5.1): if $\pi < 1 - u$ then $l_2 \to -\infty$ for $\mu_2 - \mu_1 \to 0$. Thus for a small difference in means all observations will be assigned to component 2 with high probability. With an increasing difference at first the components of some observations become uncertain before the difference gets large enough so that some observations are also assigned with high probability to component 1. For small π and a small difference in means the first condition for cluster analysis is thus met very well while the second is violated. The two conditions small q and high probability assignment of observations to both groups seem to be reasonably well satisfied for all values of π if $\mu_2 - \mu_1 > 3$.

The above demonstrates that the components of a mixture need to be sufficiently separated if the model is to be used for cluster analysis. An alternative is to combine components that are close together to one "super-component". As in Chapter 3 the distribution in this super-component is then flexibly modeled by the mixture of the component distributions. A cluster that has a skewed distribution could, for example, be fitted by several components in a mixture of normals. Leisch (2004) suggests combining components that have a small Kullback-Leibler divergence (3.2). Ray and Lindsay (2005) consider the number of modes of a multivariate mixture of normals. They show that any mode (as well as each local minimum and saddle point) has to lie in a so-called *ridgeline surface*, which for $K = 2$ reduces to the *ridgeline* from μ_1 to μ_2 given by

$$y(\gamma) = \left[(1-\gamma)\Sigma_1^{-1} + \gamma\Sigma_2^{-1}\right]^{-1} \left[(1-\gamma)\Sigma_1^{-1}\mu_1 + \gamma\Sigma_2^{-1}\mu_2\right] \quad ,$$

where $\gamma \in [0, 1]$. While the ridgeline itself does not depend on π, the location of modes on the ridgeline does. Based on these results Ray and Lindsay (2005) give detailed conditions for a mixture of two normals to be unimodal or to have more than one mode. One of their surprising findings is that such

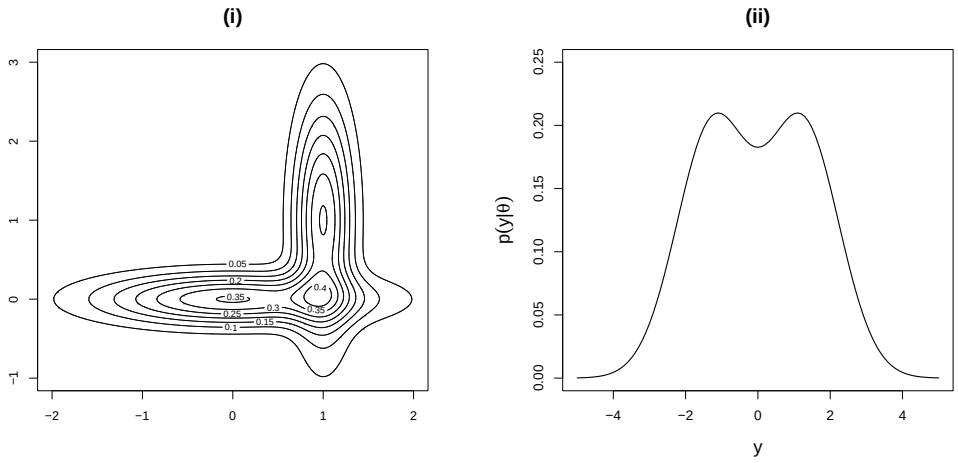

Figure 5.2: (i) Contour plot of a mixture of two bivariate normal distributions that has three modes and (ii) density of a mixture of two univariate normals that has (barely) two modes.

a mixture can have more than two modes, see Figure 5.2 (i) for an example. More than two modes can, however, not occur if the covariance matrices are proportional, i.e., $\Sigma_1 = \omega\Sigma_2$ with $\omega > 0$. From this it follows that a univariate mixture of normals can have no more than K modes, as $\sigma_k^2 = \omega\sigma_{k'}^2$ is always fulfilled. Ray and Lindsay (2005) suggest combining components that form a single mode. In a subsequent paper Li, Ray, and Lindsay (2007) extend this to components that form separate modes but where there is only a small "dip" in the ridgeline between the modes, see Figure 5.2 (ii) for an example.

5.2 Priors Induced by Bayesian Mixtures

When assigning a prior to the parameters of a Bayesian mixture model one implicitly assigns priors to other relevant quantities as well, such as a prior on the clusterings itself. Also of interest are the induced prior on the number of components K and on the pairwise clustering probability, i.e., the probability

$P(Z_i = Z_j)$ that two observations i and j share a cluster. The induced priors are given for several mixture models in the next subsections. Section 5.2.4 then discusses how these priors can help in the choice of a hyperprior for the parameter α of the Dirichlet process.

5.2.1 Priors on Clusterings

The number of possible clusterings of a set of n objects is known as the nth Bell number B_n (Bell, 1934; Rota, 1964). With $B_0 = B_1 = 1$ the Bell numbers can be computed by the recurrence formula

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \ .$$

The formula holds because the observation $n+1$ can be added as a singleton clustering to each of the B_n clusterings, it can form a cluster of two with each of the n observations which can be combined with any of the B_{n-1} clusterings of the remaining $n-1$ observations, it can join a cluster of three in $\binom{n}{n-2}$ ways which can be combined with B_{n-2} other clusterings and so on. From this construction it becomes clear that the numbers get huge quickly. While B_{10} is 115975, B_{20} is already about $5.17 \cdot 10^{13}$. Computation of a posterior for all clusterings will therefore usually not be possible.

A finite mixture with K components assigns prior mass only to clusterings with K or less than K components (some groups might be left empty). For a symmetric $Dir(\boldsymbol{\alpha})$ distribution it is given by

$$P_{FM}(\mathcal{C}|\alpha) = \frac{K!}{K_{emp}!} \frac{\Gamma(K\alpha) \prod_{k=1}^{K} \Gamma(\alpha + n_k)}{\Gamma(K\alpha + n)\Gamma(\alpha)^K} \ ,$$

where K_{emp} is the number of empty components. Note that except for the factor $K!/K_{emp}!$ the prior on clusterings $\mathcal{C}$ is the same as the prior on allocation vectors $\mathbf{Z}$ given in (2.9). The factor gives the number of allocation vectors that define the same clustering.

For the Dirichlet process the prior on a clustering can be obtained from the product of the distributions of allocation variables $Z_i|\mathbf{Z}_{i-1}$ given in (2.19) (which does not depend on the labeling of $\mathbf{Z}_{i-1}$) and turns out to be

$$P_{DP}(\mathcal{C}|\alpha) = \frac{\prod_{k=1}^{K} \alpha\Gamma(n_k)}{\prod_{i=1}^{n}(\alpha+i-1)} = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{k=1}^{K} \alpha\Gamma(n_k) \ . \tag{5.2}$$

The product $\prod_{k=1}^{K} \Gamma(n_k)$ is larger for clusterings with unequal cluster sizes so that the Dirichlet process places more mass on these clusterings. This is not surprising considering the derivation of the Dirichlet process as limit of a finite mixture model with a $Dir(\alpha/K^*, \ldots, \alpha/K^*)$ prior on $\boldsymbol{\pi}$ discussed in Subsection 2.2.2, as this prior places most mass on $\boldsymbol{\pi}$'s with rather different π_k. The influence of the favoring of unequal cluster sizes on posterior inference is discussed by Green and Richardson (2001).

For the Pitman-Yor process the prior is given by

$$P_{PY}(\mathcal{C}|\alpha,\beta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{k=1}^{K} \left[(\alpha+(k-1)\beta) \frac{\Gamma(n_k-\beta)}{\Gamma(1-\beta)} \right] \ , \tag{5.3}$$

see Pitman (2006, p. 61). As the Pitman-Yor process reduces to the Dirichlet process for $\beta = 0$, (5.3) behaves similar to (5.2) for β close to 0. For larger values of β the Pitman-Yor process tends to assign more mass to clusterings with many singletons. For $\beta \to 1$ all mass is given to the clustering where each observation is in its own cluster. It should be noted that a clustering with a small number of about equally sized clusters does not have a high prior probability under either the Dirichlet or the Pitman-Yor process.

Pitman (2006) also considers the important concept of an exchangeable prior on clusterings, which, for example, implies that the prior behaves consistently when an additional observation is added, i.e.,

$$P(\mathcal{C}) = \sum_{k=1}^{K+1} P(\mathcal{C}, Z_{n+1} = k) \ . \tag{5.4}$$

The priors discussed above can all be shown to be exchangeable. A prior that is not exchangeable is considered by Jensen and Liu (2008). It is defined by changing the transition probability of joining an existing cluster in the Pólya urn scheme (2.19) of the Dirichlet process from $n_k/(\alpha + i - 1)$ to $1/(\alpha + i - 1)$, leading to a prior that puts more weight on clusterings with about equally sized clusters. See Jensen and Liu (2008) for possibilities to deal with the non-exchangeability. Based on ideas of Hartigan (1990), Quintana and Iglesias (2003) directly assign a prior to the space of all clusterings without defining an underlying mixture model, which is then referred to as a product partition model. Because of computational simplicity the prior they mostly use is, however, (5.2), i.e., the one arising from the Dirichlet process. Product partition methods provide, however, the possibility to use more general priors on clusterings.

5.2.2 Priors on Number of Clusters

For a finite mixture model one can assign a prior on the number of clusters K, which is usually taken either to be a Poisson distribution or as a uniform distribution on $\{1, \ldots, K_{max}\}$. For infinite mixture models only a finite number of groups will be observed in a sample, so that they also induce a prior on the number of clusters. This prior will however depend on n. For the Dirichlet process Antoniak (1974) derives the expression

$$P_{DP}(K = k|\alpha, n) = \frac{\left[{n \atop k}\right]\alpha^k}{\sum_{i=1}^{n}\left[{n \atop i}\right]\alpha^i} = \frac{\Gamma(\alpha)\left[{n \atop k}\right]\alpha^k}{\Gamma(\alpha + n)} \quad , \tag{5.5}$$

where $\left[{n \atop k}\right]$ denotes a Stirling number of the first kind. Properties of Stirling numbers are discussed in Graham et al. (1989, Chapter 6), it holds, for example, that $\sum_{i=1}^{n}\left[{n \atop i}\right]\alpha^i = \prod_{i=1}^{n}(\alpha+i-1)$, justifying the second equality in (5.5). The expected number of clusters can be derived by considering the Pólya urn

scheme (2.19) and noting that the random variables W_i, with $W_i = 1$ if observation i starts a new cluster and $W_i = 0$ otherwise, have independent Bernoulli distributions with parameters $\alpha/(\alpha + i - 1)$. It follows that

$$E_{DP}(K|\alpha, n) = E\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} E(W_i) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \ .$$

Because of the relation between the logarithm and the harmonic series the expected number of clusters for the Dirichlet process behaves asymptotically as $\alpha \log n$. For the Pitman-Yor process the distribution of the number of clusters is given by a relatively complicated expression (see Pitman, 2006, p. 65) and the expected value is found as

$$E_{PY}(K|\alpha, \beta, n) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + \beta)} \frac{\Gamma(\alpha + 1)}{\beta\Gamma(\alpha + n)} - \frac{\alpha}{\beta} \ .$$

The asymptotic behavior is as $\frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha+\beta)} n^{\beta}$, so that the number of clusters grows much faster with the sample size than for the Dirichlet process.

5.2.3 Priors on Pairwise Clustering Probabilities

Model	FMfix	FMsymDir	DPM	PYPM
$P(Z_i = Z_j)$	$\sum_{k=1}^{K} \pi_k^2$	$\frac{\alpha+1}{K\alpha+1}$	$\frac{1}{\alpha+1}$	$\frac{1-\beta}{\alpha+1}$

Table 5.1: Pairwise prior clustering probabilities for several mixture models. FMfix: Finite mixture model with fixed $\boldsymbol{\pi}$, FMsymDir: Finite mixture model with symmetric $Dir(\alpha)$ prior on $\boldsymbol{\pi}$, DPM: Dirichlet process mixture, PYPM: Pitman-Yor process mixture.

Use of the posterior probabilities $P(Z_i = Z_j|\mathbf{y})$ that observation i and j are in one cluster for inference is discussed in Section 5.4. Table 5.1 gives the prior probabilities $P(Z_i = Z_j)$ for several mixture models. As mentioned above these models induce exchangeable priors on all clusterings and thus fulfill equation (5.4). The expressions in Table 5.1 can therefore be derived

by considering the clustering of only two observations. For the finite mixture model with fixed $\boldsymbol{\pi}$ the probability is given by $\sum_{k=1}^{K} \pi_k^2$, the probability of drawing two consecutive observations from the same component. For the other models it is possible to show that the entries in the table are equal to $E(\sum_{k=1}^{K} \pi_k^2)$, respectively $E(\sum_{k=1}^{\infty} \pi_k^2)$. For example, Ongaro and Cattaneo (2004) show for the Dirichlet process that, if $G \sim DP(\alpha, G_0)$, the variance of $G(A)$, with $A \in \mathcal{A}$, can be written as

$$Var(G(A)) = E\left(\sum_{k=1}^{\infty} \pi_k^2\right) G_0(A)(1 - G_0(A)) \ .$$

By comparison with (2.16) one obtains

$$E\left(\sum_{k=1}^{\infty} \pi_k^2\right) = \frac{1}{\alpha + 1} \ .$$

5.2.4 Prior Setting in Dirichlet Process Mixture Models

It has been seen in the previous subsections that the prior clustering behavior of the Dirichlet process depends only on the parameter α. It is therefore advisable to place a hyperprior on this parameter in a cluster analysis application. Escobar and West (1995) show that the conditional distribution $p(\alpha|\mathbf{Z}, \mathbf{y})$ does only depend on the number of clusters K in $\mathbf{Z}$. The distribution

$$p(\alpha|k) \propto p(\alpha)P(K = k|\alpha)$$

can then be used for sampling α in an MCMC algorithm, where $P(k|\alpha)$ is given in (5.5). They also show that if the prior on α is chosen as a Gamma distribution, $p(\alpha|k)$ can be written as a mixture of two Gammas, which makes Gibbs sampling possible. Griffin and Steel (2006) use an inverted Beta distribution as $p(\alpha)$ which is parameterized by a median n_0 and a parameter

η controlling the variability. Medvedovic et al. (2004) propose to use an inverted Gamma distribution with parameters 0.5 and 0.5 as an uninformative prior.

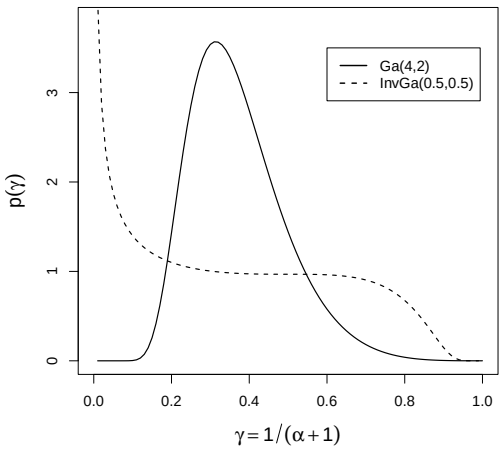

Figure 5.3: Priors induced by two different choices of $p(\alpha)$ on $\gamma = 1/(\alpha + 1)$.

The choice of the prior is usually guided by considering the induced prior on the number of components K. In Section 5.4 it is proposed to base posterior inference on $\pi_{ij} = P(Z_i = Z_j|\mathbf{y})$. In that case it is more important to consider the induced prior on $\gamma = P(Z_i = Z_j)$, which for the Dirichlet process is given by $1/(\alpha + 1)$ (see Table 5.1). Figure 5.3 shows this induced prior for two different choices of $p(\alpha)$. It can be seen that the $InvGa(0.5, 0.5)$ prior of Medvedovic et al. (2004) places a lot of weight on γ being near 0. This could lead to many observations being put into singleton clusterings and is in this context not uninformative at all. When there is a cluster structure present in the data most observations will be in a cluster with only some of the other observations. A prior that places most mass on $P(Z_i = Z_j)$ being between 0.1 and 0.6, like the $Ga(4, 2)$ also shown in Figure 5.3, seems a reasonable

default choice if a cluster structure is expected.

Alternatively, a prior could be assigned directly to $1/(\alpha+1)$. Using a $Beta(v_1, v_2)$ for this in turn induces a prior on α given by

$$p(\alpha) = \frac{\Gamma(v_1+v_2)}{\Gamma(v_1)\Gamma(v_2)}\alpha^{v_2-1}(\alpha+1)^{-(v_1+v_2)} \quad . \tag{5.6}$$

This is a Beta distribution of the second kind (Johnson et al., 1995, p. 248). A possible choice for an uninformative prior is a uniform (i.e., $Beta(1,1)$) prior on the prior clustering probability leading to $p(\alpha) = 1/(\alpha+1)^2$.

5.3 Clustering With a Fixed Number of Clusters

5.3.1 Label-Switching

It has been discussed in Subsection 2.1.3 that the likelihood of a finite mixture model is not identifiable. One reason for this is that a permutation of the component labels $\{1, \ldots, K\}$ does not change the likelihood, which has thus $K!$ modes of equal height. With exchangeable priors $p(\theta_k)$ and $p(\pi_k)$ the same holds for the posterior distribution. This is not a concern if one wants to obtain a clustering $\mathcal{C}$ of the observations via an allocation vector $\widehat{\mathbf{Z}}$ that maximizes the posterior $p(\mathcal{C}(\mathbf{Z})|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{Z})p(\mathcal{C}(\mathbf{Z}))$ since the allocation vectors of all modes define the same clustering $\mathcal{C}$. However, when using an MCMC algorithm to sample from the posterior, it is likely that more than one mode is visited during the run, a phenomenon called "label-switching". This will necessarily occur if the MCMC sampler is run long enough, as it is guaranteed to eventually visit all the modes. An illustration of label-switching is given in Figure 5.4. It shows the mean of the second component during the MCMC fitting of one of the simulated data sets described in Subsection 5.5.1. Depending on which of the real clusters the component represents

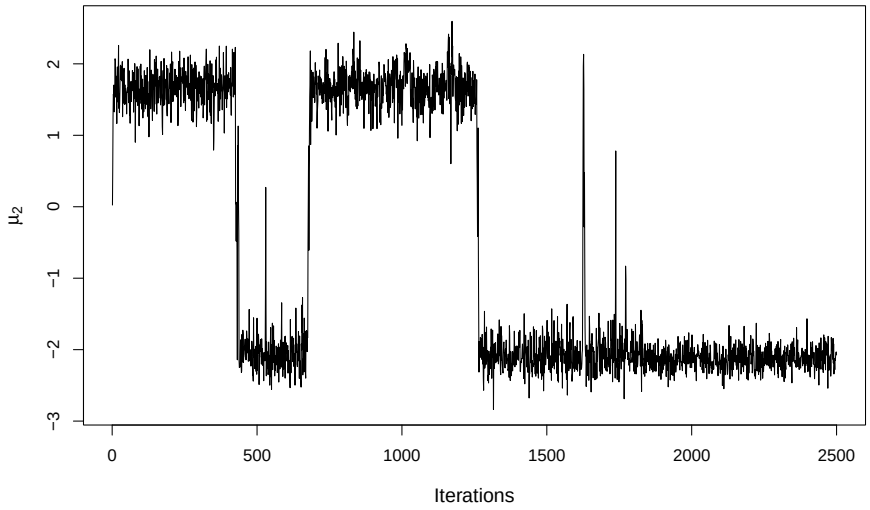

Figure 5.4: Illustration of label-switching.

the values are either around -2 or around 2. So when one has obtained an MCMC sample of allocation vectors $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$ the intuitive estimator for the posterior probability that observation i belongs to cluster k,

$$P(Z_i = k|\mathbf{y}) \approx \frac{1}{M} \sum_{m=1}^{M} I_{\{Z_i^{(m)}=k\}} \ , \tag{5.7}$$

with $i = 1, \ldots, n$ and $k = 1, \ldots, K$, does not lead to sensible results when applied to the unprocessed sample. In (5.7) the running index m is used instead of the index t employed in Chapter 2. This is to indicate that possibly some processing like discarding a burn-in and thinning has taken place.

5.3.2 Identifiability Constraints

One approach to deal with label-switching to impose an identifiability constraint that is fulfilled by only one of the $K!$ permutations, e.g., imposing $\mu_1 < \mu_2 < \ldots < \mu_K$ in a univariate mixture of normals. This restricts the

prior $p(\boldsymbol{\theta}, \boldsymbol{\pi})$, and thus the posterior, to a region containing only one mode. Stephens (1997, pp. 43-44) showed that it is not necessary to include the constraint in the MCMC run, which might be difficult or inefficient. Running an unconstrained sampler and post-processing the sample by permuting each $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\pi}^{(m)})$ such that the constraint is fulfilled also leads to a sample from the constrained posterior. It is, however, not the case that an arbitrary identifiability constraint induces a unique labeling. Although the constrained region contains only one mode there might still be considerable posterior mass from other modes in it. This happens if some of the constraints are only barely fulfilled, e.g., if μ_1 is only a bit smaller than μ_2 in a univariate mixture of normals. So especially for a higher-dimensional $\boldsymbol{\theta}$ it can be difficult to find a suitable constraint. See Stephens (2000) and Frühwirth-Schnatter (2006, pp. 44-49) for additional discussion and examples.

5.3.3 Relabeling Algorithms

Stephens (2000) proposes to use so-called relabeling algorithms instead of identifiability constraints. These are based on a loss function $L(\mathbf{P}, (\boldsymbol{\theta}, \boldsymbol{\pi}))$, where $\mathbf{P}$ is the $n \times K$ matrix of allocation probabilities $P(Z_i = k) \doteqdot p_{ik}$. The basic loss function proposed by Stephens is given by

$$L_0^S(\mathbf{P}, (\boldsymbol{\theta}, \boldsymbol{\pi})) = \sum_{i=1}^{n} \sum_{k=1}^{K} P(Z_i = k|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\pi}) \log \frac{P(Z_i = k|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\pi})}{p_{ik}} \ , \qquad (5.8)$$

where $P(Z_i = k|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{\pi_k p(y_i|\theta_k)}{\sum_{l=1}^{K} \pi_l p(y_i|\theta_l)}$ and the p_{ik} need to fulfill the constraints $\sum_k p_{ik} = 1$, $i = 1, \ldots, n$. Here we propose the use of a variant of the loss function of Stephens (2000) that depends on $\mathbf{Z}$ rather than $(\boldsymbol{\theta}, \boldsymbol{\pi})$. We will first introduce the resulting algorithm and then compare it with Stephens' approach.

The basic loss function of our approach is given by

$$L_0(\mathbf{P}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} -\log p_{ik} \cdot I_{\{Z_i=k\}} \ . \tag{5.9}$$

Using the convention $\log 0 \cdot 0 = 0$ the loss is zero if $p_{ik} = 1$ whenever $Z_i = k$ and the loss will be bigger than zero otherwise. Since a permutation ν of the cluster labels does not change the clustering, the loss function should be invariant to such a permutation so that the loss function actually employed will be

$$L(\mathbf{P}, \mathbf{Z}) = \min_{\nu} L_0(\mathbf{P}, \nu(\mathbf{Z})) = \min_{\nu} \sum_{i=1}^{n} \sum_{k=1}^{K} -\log p_{ik} \cdot I_{\{\nu(Z_i)=k\}} \ .$$

A decision-theoretic approach is now to choose $\widehat{\mathbf{P}}$ such that the posterior expected loss (or posterior risk) $E(L(\mathbf{P}, \mathbf{Z})|\mathbf{y})$ is minimized at $\widehat{\mathbf{P}}$. It is not feasible to compute the posterior risk directly since this involves summation over the K^n possible values of $\mathbf{Z}$ but it can be approximated from the MCMC sample $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$ by

$$E(L(\mathbf{P}, \mathbf{Z})|\mathbf{y}) \approx \frac{1}{M} \sum_{m=1}^{M} \min_{\nu_m} \sum_{i=1}^{n} \sum_{k=1}^{K} -\log p_{ik} \cdot I_{\{\nu_m\left(Z_i^{(m)}\right)=k\}} \ . \tag{5.10}$$

The following algorithm can then be used to find a $\widehat{\mathbf{P}}$ minimizing (5.10).

Algorithm 5.1: Allocation Relabeling:

Start with some initial values for $\nu_1, \ldots, \nu_M$ and repeat until convergence:

1) Set $\widehat{p}_{ik} = \frac{1}{M} \sum_{m=1}^{M} I_{\{\nu_m\left(Z_i^{(m)}\right)=k\}}$.

2) For $m = 1, \ldots, M$: Choose ν_m to minimize

$$\sum_{i=1}^{n} \sum_{k=1}^{K} -\log \widehat{p}_{ik} \cdot I_{\{\nu_m\left(Z_i^{(m)}\right)=k\}} \ .$$

Steps 1 and 2 decrease (5.10) (this is obvious for step 2, see Lemma 5.1 below for step 1). Repeating both steps is then guaranteed to reach a local minimum of (5.10), as there are only finitely many permutations $\nu_1, \ldots, \nu_M$. It is, however, not guaranteed that a global minimum is found, so that several starting values for the labels should be tried.

Lemma 5.1: *The value of p_{ik} that minimizes (5.10), subject to the constraint $\sum_k p_{ik} = 1$, is given by $\widehat{p}_{ik} = \frac{1}{M} \sum_{m=1}^{M} I_{\{\nu_m\left(Z_i^{(m)}\right)=k\}}$.*

Proof: Differentiating (5.10) with respect to p_{ik} and including the constraint $\sum_k p_{ik} = 1$ with a Lagrange multiplier leads to

$$\begin{aligned} & \frac{\partial}{\partial p_{ik}} \left[\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{k=1}^{K} - \log p_{ik} \cdot I_{\{Z_i^{(m)}=k\}} + \lambda \left(\sum_{k=1}^{K} p_{ik} - 1 \right) \right] \\ = & \frac{1}{M} \sum_{m=1}^{M} - \frac{\partial}{\partial p_{ik}} \log p_{ik} \cdot I_{\{Z_i^{(m)}=k\}} + \lambda \\ = & -\frac{1}{M p_{ik}} \sum_{m=1}^{M} I_{\{Z_i^{(m)}=k\}} + \lambda \end{aligned}$$

Setting the last expression to 0 leads to

$$p_{ik} = \frac{1}{M\lambda} \sum_{m=1}^{M} I_{\{Z_i^{(m)}=k\}} \ ,$$

which, using $\sum_k p_{ik} = 1$, is solved by $\lambda = 1$ and thus $\widehat{p}_{ik} = \frac{1}{M} \sum_{m=1}^{M} I_{\{Z_i^{(m)}=k\}}$. By taking the second derivative it is easily verified that this is indeed a minimum. $\square$

It is not necessary to try all possible permutations ν_m in step 2, as the minimization can be formulated as an instance of the *assignment problem* of linear programming, see Stephens (2000) for details. For this problem efficient algorithms with running times of the order $\mathcal{O}(K^3)$ are available.

The algorithm originally proposed by Stephens (based on the basic loss (5.8)) involves the minimization of an expression that is analog to (5.10). By comparing (5.8) and (5.9) it can be seen that the algorithm proposed here effectively replaces $P(Z_i = k|\mathbf{y}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\pi}^{(m)})$ by $I_{\{Z_i^{(m)}=k\}}$. Although the former possibly contains a bit more information about the group structure, using the $\mathbf{Z}^{(m)}$ for relabeling offers two advantages. First it is computationally less demanding. Stephens' algorithm requires the computation and storage of M matrices of dimension $n \times K$ containing the probabilities $P(Z_i = k|\mathbf{y}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\pi}^{(m)})$ which is not necessary for Algorithm 5.1. The second advantage is that the algorithm does not depend on the form of the component densities $p(y|\theta)$. It is therefore more general and can be applied without modification to the MCMC output of any finite mixture model. It can also be used in connection with samplers that only sample $\mathbf{Z}$, e.g., the one of Chen and Liu (1996) discussed in Subsection 2.1.4.

5.4 Clustering With a Varying Number of Clusters

When an infinite mixture or a finite mixture with a prior on K is fitted via an MCMC algorithm an additional complication in summarizing the sample of allocation vectors $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$ for clustering inference arises. Now not only the labeling of groups can differ between the $\mathbf{Z}^{(m)}$ but also the number of groups there are. The relabeling algorithm of the last section can no longer be applied. A possible solution proposed by Tadesse et al. (2006) is to apply the algorithm only to those $\mathbf{Z}^{(m)}$, where the number of groups $K^{(m)}$ is equal to $\widehat{K} = \arg\max_k \sum_{m=1}^{M} I_{\{K^{(m)}=k\}}$, i.e., the most commonly occurring number of groups. This means first obtaining a MAP estimate of K and then doing inference conditional on that estimate. The approach might, however, be

wasteful if the marginal posterior $P(K|\mathbf{y})$ does not have a pronounced mode. Another possibility is basing the inference on the pairwise posterior clustering probability, which will be considered in the following.

5.4.1 The Posterior Similarity Matrix

The matrix of pairwise posterior probabilities that observations i and j are in one cluster $P(Z_i = Z_j|\mathbf{y})$ will be denoted as posterior similarity matrix (PSM). It is easy to see that label-switching does not affect the PSM which can therefore be estimated from the MCMC sample by

$$\pi_{ij} = P(Z_i = Z_j|\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M} I_{\{Z_i^{(m)}=Z_j^{(m)}\}} \ . \tag{5.11}$$

In (5.11) the number of groups does not have to be fixed.

A distance matrix is given by 1-PSM, which contains the posterior probabilities that the observations i and j are not in one cluster. These probabilities define a sensible distance measure, as shown in the following lemma.

Lemma 5.2: *$1-\pi_{ij}$ is a topological pseudometric for the space of observations as it fulfills the conditions*

$$1-\pi_{ii} = 0 \tag{5.12}$$

$$1-\pi_{ij} = 1-\pi_{ji} \tag{5.13}$$

$$\text{and} \quad (1-\pi_{ij}) \leq (1-\pi_{il}) + (1-\pi_{jl}) \ , \tag{5.14}$$

for all $i, j \in \{1, \ldots, n\}$.

Proof: (5.12) and (5.13) are straightforward to see. $1-\pi_{ij}$ is not a metric since $1-\pi_{ij} = 0$ does not imply that the observations i and j are equal. It remains to show that the triangle inequality (5.14) is valid, which is equivalent

to

$$\begin{aligned} & (1-\pi_{ij}) \leq (1-\pi_{il}) + (1-\pi_{jl}) \\ \Longleftrightarrow \quad & \pi_{ij} \geq \pi_{il} + \pi_{jl} - 1 \ . \end{aligned}$$

In every possible clustering the observations i,j and l are grouped according to one of the patterns

$$\begin{aligned} &\text{I} : \{i,j,l\} && \text{II} : \{i,j\},\{l\} && \text{III} : \{i\},\{j,l\} \\ &\text{IV} : \{i,l\},\{j\} && \text{V} : \{i\},\{j\},\{l\} \ . \end{aligned}$$

Then the following equations hold

$$\pi_{ij} = P(\text{I}|\mathbf{y}) + P(\text{II}|\mathbf{y}) \tag{5.15}$$

$$\pi_{jl} = P(\text{I}|\mathbf{y}) + P(\text{III}|\mathbf{y}) \tag{5.16}$$

$$\pi_{il} = P(\text{I}|\mathbf{y}) + P(\text{IV}|\mathbf{y}) \tag{5.17}$$

$$1 = P(\text{I}|\mathbf{y}) + \cdots + P(\text{V}|\mathbf{y}) \ , \tag{5.18}$$

and with (5.16) and (5.17) one obtains

$$\begin{aligned} \pi_{jl} + \pi_{il} - 1 &= 2 \cdot P(\text{I}|\mathbf{y}) + P(\text{III}|\mathbf{y}) + P(\text{IV}|\mathbf{y}) - 1 \\ &\overset{(5.15)}{=} P(\text{I}|\mathbf{y}) + P(\text{III}|\mathbf{y}) + P(\text{IV}|\mathbf{y}) - P(\text{II}|\mathbf{y}) + \pi_{ij} - 1 \\ &\overset{(5.18)}{=} -2 \cdot P(\text{II}|\mathbf{y}) - P(\text{V}|\mathbf{y}) + \pi_{ij} \\ &\leq \pi_{ij} \qquad \square \end{aligned}$$

The posterior similarity matrix can be displayed as a heatmap to give an overview over the clustering of observations. An example involving the iris data, which will be discussed in detail in Subsection 5.5.4, is given in Figure 5.5 (i). Since 1-PSM is a distance matrix another possibility for visualization is to apply multidimensional scaling (see, e.g., Cox and Cox, 2001) to 1-PSM, giving a set of points that have relative Euclidean distances which

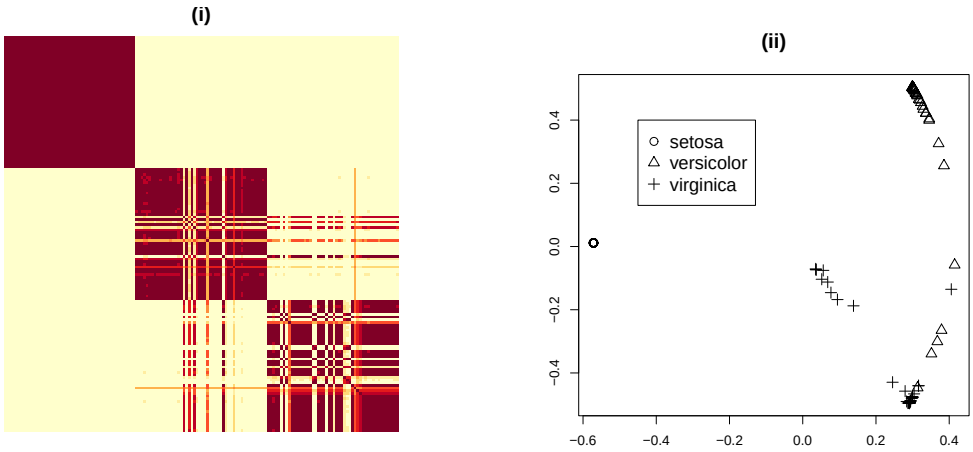

Figure 5.5: (i) Heatmap of PSM for iris data. Light yellow indicates low π_{ij}, dark red high π_{ij}. (ii) Classical multidimensional scaling applied to PSM of iris data.

closely match the distances $1 - \pi_{ij}$. See Figure 5.5 (ii) for an example with the iris data.

5.4.2 Clustering Methods Based on the Posterior Similarity Matrix

The sample $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$ contains a lot of information, which is however hard to overlook in its entirety. The PSM already provides a useful summary. Usually one would also like to summarize the sample with a single allocation vector $\widehat{\mathbf{Z}}$, to obtain a summarizing clustering $\mathcal{C}(\widehat{\mathbf{Z}})$. We now consider methods for choosing $\widehat{\mathbf{Z}}$ based on the PSM.

Ad hoc approach of Medvedovic et al. (2004)

Medvedovic et al. (2004) employ agglomerative hierarchical clustering, as described in Subsection 4.1.2, to obtain an estimate $\widehat{\mathbf{Z}}$, using $1 - \pi_{ij}$ as the distance between the observations i and j. If K is known they use average linkage and cut the dendrogram at K groups. For unknown K they use complete linkage and cut the dendrogram at a distance of $1 - \varepsilon$, for small,

positive ε. For two distinct clusters C_k and $C_{k'}$ there is then at least one pair of observations (i, j) with $Z_i = k$ and $Z_j = k'$, such that $\pi_{ij} < \varepsilon$. In the applications in the next section a value of $\varepsilon = 0.01$ will be employed.

Binder's loss function

Binder (1978) was the first to consider loss functions based on pairwise occurrences of observations, i.e.,

$$L(\mathbf{Z}^*, \mathbf{Z}) = \sum_{i<j} \ell_1 \cdot I_{\{Z_i^* \neq Z_j^*\}} I_{\{Z_i = Z_j\}} + \ell_2 \cdot I_{\{Z_i^* = Z_j^*\}} I_{\{Z_i \neq Z_j\}} \ , \tag{5.19}$$

with positive constants ℓ_1 and ℓ_2. $\mathbf{Z}^*$ is a proposed estimate and the matrix containing $I_{\{Z_i^* = Z_j^*\}}$ is a known 0-1 matrix, which will be referred to as estimated similarity matrix. The unknown true allocation vector $\mathbf{Z}$ has a similarity matrix containing $I_{\{Z_i = Z_j\}}$. Since $E(I_{\{Z_i = Z_j\}}|\mathbf{y}) = \pi_{ij}$, the posterior similarity matrix can be seen as the similarity matrix of the posterior expected clustering $E(\mathcal{C}(\mathbf{Z})|\mathbf{y})$.

The two parts of (5.19) represent two different goals of cluster analysis, *completeness* and *homogeneity*. Completeness means that all observations from a true group are together in an estimated cluster and lack of it is penalized by the term $\ell_1 \cdot I_{\{Z_i^* \neq Z_j^*\}} I_{\{Z_i = Z_j\}}$. Homogeneity is achieved if each estimated cluster contains only observations from one true group. Lack of it is penalized by $\ell_2 \cdot I_{\{Z_i^* = Z_j^*\}} I_{\{Z_i \neq Z_j\}}$. There is a trade-off between the two goals, similarly to the trade-off between bias and variance in point estimation. Completeness can be better achieved by relatively large estimated clusters whereas relatively small estimated clusters are better for homogeneity. One of the goals can be perfectly achieved while sacrificing the other by either putting all observations into one large cluster or all observations into singleton clusters. The quotient ℓ_1/ℓ_2 determines the relative importance of the two goals. When there is no particular preference a pragmatic solution is to

set $\ell_1 = \ell_2 = 1$ and thus to weight the goals equally. This approach is taken by Hurn et al. (2003) in the context of a switching regression model. In the following we will refer to the $\ell_1 = \ell_2 = 1$ case of (5.19) as Binder's loss.

The posterior expectation of this loss can be written as

$$E(L(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y}) = \sum_{i<j} |I_{\{Z_i^* = Z_j^*\}} - \pi_{ij}| \ , \tag{5.20}$$

i.e., the sum of absolute deviations of the estimated similarity matrix to the posterior similarity matrix. The estimated $\widehat{\mathbf{Z}}$ can be taken as the allocation vector $\mathbf{Z}^*$ minimizing (5.20). Because of the linearity of the loss function the same expression is obtained if the loss function is computed between estimated and posterior expected clustering, so that

$$E(L(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y}) = L(\mathbf{Z}^*, E(\mathcal{C}(\mathbf{Z})|\mathbf{y})) \ . \tag{5.21}$$

With Binder's loss function a loss of 1 is made whenever a pair of observations is treated differently in the estimated clustering $\mathcal{C}(\mathbf{Z}^*)$ than in the true $\mathcal{C}(\mathbf{Z})$. The loss is thus the sum of disagreements in the treatment of pairs of observations between the estimated and true clustering. Considering the Rand index of Section 4.2 it can then be seen that

$$R(\mathcal{C}(\mathbf{Z}^*), \mathcal{C}(\mathbf{Z})) = 1 - \frac{L(\mathbf{Z}^*, \mathbf{Z})}{\binom{n}{2}} \ .$$

The $\widehat{\mathbf{Z}}$ that minimizes the posterior expectation of Binder's loss in equation (5.20) also maximizes the posterior expected Rand index with the true clustering and, considering equation (5.21), the Rand index of estimated and posterior expected clustering. For simplicity we will write $R(\mathcal{C}(\mathbf{Z}^*), \mathcal{C}(\mathbf{Z}))$ as $R(\mathbf{Z}^*, \mathbf{Z})$ in the following.

Dahl's criterion

Dahl (2006) proposed

$$\sum_{i<j} (I_{\{Z_i^*=Z_j^*\}} - \pi_{ij})^2 \tag{5.22}$$

as a heuristic criterion to be minimized to obtain an estimate $\widehat{\mathbf{Z}}$, without giving a decision-theoretic motivation. It turns out that minimization of (5.20) and (5.22) is equivalent, which can be seen by writing

$$\sum_{i<j} |I_{\{Z_i^*=Z_j^*\}} - \pi_{ij}| = \sum_{i<j} (\pi_{ij} - 2 \cdot I_{\{Z_i^*=Z_j^*\}}\pi_{ij} + I_{\{Z_i^*=Z_j^*\}}) \ ,$$

and

$$\sum_{i<j} (I_{\{Z_i^*=Z_j^*\}} - \pi_{ij})^2 = \sum_{i<j} (\pi_{ij}^2 - 2 \cdot I_{\{Z_i^*=Z_j^*\}}\pi_{ij} + I_{\{Z_i^*=Z_j^*\}}) \ .$$

The difference between these sums is $\sum_{i<j} \pi_{ij}(1 - \pi_{ij})$, which does not depend on $\mathbf{Z}^*$, so that minimization of Dahl's criterion is equivalent to the minimization of the posterior expectation of Binder's loss.

Posterior Expected Adjusted Rand

As discussed in Section 4.2 the adjusted Rand index is usually preferable as a measure of association to the unadjusted index. It is also used by Dahl (2006) and Medvedovic et al. (2004) in the evaluation of their simulation studies. Fritsch and Ickstadt (2009) therefore propose to maximize the adjusted Rand index of estimated and true clustering $AR(\mathbf{Z}^*, \mathbf{Z})$ instead of $R(\mathbf{Z}^*, \mathbf{Z})$, as is done with the minimization of Binder's loss.

Suppose that the estimated and true clustering corresponds to clustering

$\mathcal{C}$ and $\mathcal{C}'$ in Table 4.1, respectively. In that case the equations

$$\begin{aligned}\textstyle\sum_k \binom{n_{k\cdot}}{2} &= \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \ , \\ \textstyle\sum_l \binom{n_{\cdot l}}{2} &= \sum_{i<j} I_{\{Z_i=Z_j\}} \quad \text{and} \\ \textstyle\sum_{k,l} \binom{n_{kl}}{2} &= \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} I_{\{Z_i=Z_j\}} \ ,\end{aligned}$$

hold. The adjusted Rand index of (4.5) can then be written as

$$\frac{\sum_{i<j} I_{\{Z_i^*=Z_j^*\}} I_{\{Z_i=Z_j\}} - \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \sum_{i<j} I_{\{Z_i=Z_j\}} / \binom{n}{2}}{\frac{1}{2}[\sum_{i<j} I_{\{Z_i^*=Z_j^*\}} + \sum_{i<j} I_{\{Z_i=Z_j\}}] - \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \sum_{i<j} I_{\{Z_i=Z_j\}} / \binom{n}{2}} \ .$$

This expression depends of course on the unknown true allocation $\mathbf{Z}$. When taking the posterior expectation to obtain an expression that can be computed given a potential estimate $\mathbf{Z}^*$ and the MCMC sample $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \ldots, \mathbf{Z}^{(M)}$ we can utilize either side of equation (5.21) leading to either $E(AR(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y})$ or $AR(\mathbf{Z}^*, E(\mathcal{C}(\mathbf{Z})|\mathbf{y}))$. Unlike Binder's loss $AR(\mathbf{Z}^*, \mathbf{Z})$ is not a linear function of the $I_{\{Z_i=Z_j\}}$, so these two expressions are related but not the same.

Maximizing $E(AR(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y})$ over $\mathbf{Z}^*$ leads to a clustering that has maximum posterior expected adjusted Rand index with the true clustering. This can be approximated from the MCMC sample by

$$\frac{1}{M} \sum_{m=1}^{M} AR(\mathbf{Z}^*, \mathbf{Z}^{(m)}) \ , \tag{5.23}$$

with $AR(\mathbf{Z}^*, \mathbf{Z}^{(m)})$ being computed by equation (4.5).

The adjusted Rand index with the posterior expected clustering $AR(\mathbf{Z}^*, E(\mathcal{C}(\mathbf{Z})|\mathbf{y}))$ is given by the expression

$$\frac{\sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \pi_{ij} - \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \sum_{i<j} \pi_{ij} / \binom{n}{2}}{\frac{1}{2}[\sum_{i<j} I_{\{Z_i^*=Z_j^*\}} + \sum_{i<j} \pi_{ij}] - \sum_{i<j} I_{\{Z_i^*=Z_j^*\}} \sum_{i<j} \pi_{ij} / \binom{n}{2}} \ , \tag{5.24}$$

where the π_{ij} are estimated from the MCMC sample via equation (5.11). These two slightly different criteria are introduced in this thesis since both have distinctive advantages. Expression (5.24) requires the computation of the posterior similarity matrix, but can then be evaluated a lot faster than (5.23), which is advantageous if the criterion needs to be calculated for many different $\mathbf{Z}^*$. We also found (5.24) to be more amenable to a theoretical study. Expression (5.23) on the other hand does not require the computation of the posterior similarity matrix, which can be preferable for large n, where this matrix gets too large to be stored. And one can argue that maximizing $E(AR(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y})$ is a more standard approach than maximizing $AR(\mathbf{Z}^*, E(\mathcal{C}(\mathbf{Z}))|\mathbf{y}))$. Practically, however, we found in our applications that maximization of either criterion leads to nearly identical results. For simplicity we will in the following refer to both criteria as PEAR for Posterior Expected Adjusted Rand.

A possible disadvantage of using PEAR as an optimality criterion is that the adjusted Rand index is 0 if one of the compared clusterings consists of only one cluster or of all singletons. As there will always be allocations $\mathbf{Z}^*$ leading to positive values of (5.23) or (5.24) these extreme clusterings will never be chosen as $\widehat{\mathbf{Z}}$.

A shrinkage property

It is instructive to consider the behavior of Binder's loss and PEAR, if there were no restrictions for the $I_{\{Z_i^*=Z_j^*\}}$, i.e., all $I_{\{Z_i^*=Z_j^*\}}$ could be set individually to 0 or 1 without regard of the other indicator functions. From equation (5.20) it is clear that for Binder's loss the optimal solution in that case is simply to set $I_{\{Z_i^*=Z_j^*\}} = 1$, if $\pi_{ij} \geq 0.5$. In the case of PEAR it can be seen that for the expression (5.24) if ℓ of the $I_{\{Z_i^*=Z_j^*\}}$ are 1 and the rest 0, the maximum

is attained if the $I_{\{c_i^*=c_j^*\}} = 1$ correspond to the ℓ highest π_{ij}. Denoting with $\pi_{(i)}$ the ith largest π_{ij} and letting $\sum_{i<j} \pi_{ij} / \binom{n}{2} = \bar{\pi}_{..}$, the maximum of (5.24) is then given by the maximum of

$$PEAR^*(\ell) = \frac{\sum_{i=1}^{\ell} \pi_{(i)} - \ell\, \bar{\pi}_{..}}{\frac{1}{2}\ell\,(1 - 2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} \tag{5.25}$$

for $\ell = 0, 1, \ldots, \binom{n}{2}$. Then the following lemma shows that PEAR has a shrinkage property.

Lemma 5.3: *At the value ℓ^* for which (5.25) is maximal there is a threshold t such that $\pi_{(\ell^*)} = \min\{\pi_{(i)} : \pi_{(i)} \geq t\}$ and the following relation holds:*

$$\begin{aligned} \bar{\pi}_{..} < t < 0.5 \quad & \text{if} \quad \bar{\pi}_{..} < 0.5 \\ t = 0.5 \quad & \text{if} \quad \bar{\pi}_{..} = 0.5 \\ \bar{\pi}_{..} > t > 0.5 \quad & \text{if} \quad \bar{\pi}_{..} > 0.5 \ . \end{aligned} \tag{5.26}$$

Proof: See Appendix A.

Compared to Binder's loss, where t is always 0.5, Lemma 5.3 shows that for PEAR the threshold for setting $I_{\{Z_i=Z_j\}}$=1 is shrunk towards the mean of the π_{ij}. PEAR thus adjusts to the amount of clustering found in the data. If overall only few π_{ij} are large, the conditions on putting two observations in one cluster are less strict, e.g., two observations i and j might be clustered together if π_{ij} is only 0.4. The opposite applies if overall many π_{ij} are large.

5.4.3 Optimization of Criteria

Unlike the method of Medvedovic et al. (2004) the expectation of Binder's loss and PEAR have to be optimized over several $\mathbf{Z}^*$ to obtain a clustering estimate $\widehat{\mathbf{Z}}$. The same is the case for a MAP estimate, as the EM algorithm is no longer applicable with varying K. Due to the exponential growth of the

Bell numbers B_n it is not feasible to compute the criteria for all possible clusterings, except for very small n. So a small set of candidate allocations $\mathbf{Z}^*$ that will lead to a close to optimal solution is needed. A simple solution chosen, for example, by Dahl (2006) is to take the MCMC sample $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \ldots, \mathbf{Z}^{(M)}$ as this set.

Another possibility of a small set of clusterings with potentially good values that will be considered here is given by the clusterings obtained by the hierarchical clustering approach with distances $1 - \pi_{ij}$ of Medvedovic et al.. The criteria can be computed for the clusterings on every level of the hierarchy and $\widehat{\mathbf{Z}}$ taken to be the optimal among these. For the special case of the expectation of Binder's loss and the clusterings $\mathbf{Z}^*$ from the hierarchical clustering with average linkage it is not necessary to compute the criterion for all $\mathbf{Z}^*$, as the following lemma shows.

Lemma 5.4: *Among the clusterings $\mathbf{Z}^*$ given by the levels of the hierarchical clustering with average linkage and distances $1 - \pi_{ij}$, the clustering $\widehat{\mathbf{Z}}$ minimizing $E(L(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y}) = \sum_{i<j} |I_{\{Z_i^*=Z_j^*\}} - \pi_{ij}|$ is obtained by cutting the dendrogram at a height of 0.5.*

Proof: When clusters C_k and $C_{k'}$ are merged the $I_{\{Z_i^*=Z_j^*\}}$ with $i \in C_k$ and $j \in C_{k'}$ are set from 0 to 1. The change in $E(L(\mathbf{Z}^*, \mathbf{Z})|\mathbf{y}) = \sum_{i<j} |I_{\{Z_i^*=Z_j^*\}} - \pi_{ij}|$ is then

$$\begin{aligned} & \sum_{i \in C_k} \sum_{j \in C_{k'}} (1 - \pi_{ij}) - \sum_{i \in C_k} \sum_{j \in C_{k'}} \pi_{ij} \\ = & \sum_{i \in C_k} \sum_{j \in C_{k'}} (1 - 2\pi_{ij}). \end{aligned} \tag{5.27}$$

If (5.27) is smaller than 0 the merge improves the expectation of Binder's

loss. This is equivalent to

$$\begin{aligned}
& \textstyle\sum_{i\in C_k}\sum_{j\in C_{k'}}(1-2\pi_{ij}) && < 0 \\
\iff \quad & n_k n_{k'} - 2\textstyle\sum_i\sum_j \pi_{ij} && < 0 \\
\iff \quad & \tfrac{n_k n_{k'}}{2} - \textstyle\sum_i\sum_j \pi_{ij} && < 0 \\
\iff \quad & n_k n_{k'} - \textstyle\sum_i\sum_j \pi_{ij} && < \frac{n_k n_{k'}}{2} \\
\iff \quad & \textstyle\sum_i\sum_j(1-\pi_{ij}) && < \frac{n_k n_{k'}}{2} \\
\iff \quad & \frac{\sum_{i\in C_k}\sum_{j\in C_{k'}}(1-\pi_{ij})}{n_k n_{k'}} && < \frac{1}{2}\ .
\end{aligned}$$

The left hand side of the last expression is the distance $d(C_k, C_{k'})$ given in (4.3), that is used in average linkage. Merging of clusters up to a distance of 0.5 will thus improve (5.20) and merging clusters with a larger distance will deteriorate it. □

More sophisticated optimization methods could of course be applied. For the expectation of Binder's loss Lau and Green (2007) show that it suffices to minimize the linear functional

$$\sum_{i<j} I_{\{Z_i^*=Z_j^*\}}(1-2\pi_{ij})\ .$$

Using the constraints that for all triples (i, j, l), if $I_{\{Z_i^*=Z_j^*\}} = 1$ then $I_{\{Z_i^*=Z_l^*\}} = I_{\{Z_j^*=Z_l^*\}}$, they formulate the minimization of the expected loss as a binary integer programming problem, which can be solved exactly. Bansal et al. (2004) have, however, shown that this is an NP-hard problem. Lau and Green therefore propose an approximate solution where in turn each of the n observations is allocated optimally while holding the clustering of the other $n-1$ observations fixed. Even this approximation algorithm requires to solve n binary integer programming problems with $\mathcal{O}(n)$ variables and $\mathcal{O}(n^2)$ constraints in each iteration.

In machine learning the problem of combining several clusterings is known as *consensus clustering*. The clusterings to be combined are in that case not necessarily from an MCMC sample but might, for example, also result from applying K-means with different starting values or values of K. The proposed solutions mostly try to minimize a function equivalent to the expectation of Binder's loss. Goder and Filkov (2008) give a recent overview of algorithms that have been used for this goal. These include the already discussed approaches of taking the best among the given clusterings or from a hierarchical clustering with average linkage. Goder and Filkov did, however, not realize that in the latter case the best solution is always attained by cutting at 0.5, as shown in Lemma 5.4. Other approaches considered by them are greedy search and simulated annealing algorithms. They do not consider integer programming based algorithms like the one of Lau and Green (2007). In a simulation study Goder and Filkov (2008) find that overall the best performance is given by combining the average linkage algorithm with some greedy steps on the resulting clustering.

5.5 Applications

In this chapter Bayesian mixtures are used for cluster analysis on simulated and real data, applying the methods described thus far. Computation times refer to a desktop computer with 3 GHz and 2 GB RAM. A package `mcclust` for the statistical software `R` (R Development Core Team, 2009) has been written. It implements the computation of the posterior similarity matrix, the optimization of Binder's loss and PEAR as well as the relabeling algorithm of Subsection 5.3.3. The functions of the package are described in Appendix E.

5.5.1 Simulation Study

Setup

The simulated data are 3-dimensional with 8 clusters, where the cluster means are given by the 8 possible values of $(\pm\delta, \pm\delta, \pm\delta)^T$. Observations are obtained by adding independent standard normal errors to the cluster means. As δ determines how well the clusters are separated, we use values of $\delta \in \{0.5, 1.0, 1.5, 2.0\}$ to get data sets that range from ones with largely overlapping to ones with fairly well separated clusters. One scenario with equal cluster sizes is simulated, where each cluster contains 50 observations and one with unequal sizes where half of the clusters contain 20 and the other half 80 observations. For each combination of δ and cluster size 10 data sets are generated. To investigate how the clustering methods perform in extreme cases, data sets are also generated for $\delta = 0$, so that all observations come from the same normal component. A scenario were each observation comes from its own component is simulated by setting $\delta = 2$ and extending the dimensionality of the data to 9, giving 512 observations with distinct cluster means.

Dirichlet process mixture model

The used model is similar to the one proposed by Qin (2006). It assumes

$$\begin{aligned} y_i|\mu_i, \sigma_i^2 &\sim N(\mu_i, \sigma_i^2 I_p) \\ \mu_i, \sigma_i^2|G &\sim G \\ G &\sim DP(\alpha, p(\mu, \sigma^2)) \ , \end{aligned} \tag{5.28}$$

with random probability measure G. Clustering is thus induced on common values of (μ, σ^2). The base distribution is chosen as a conjugate Normal-

Inverse Gamma

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\ &= N(0, \sigma^2 v^{-1} I_p) InvGa(a, b) \ . \end{aligned}$$

As discussed in Subsection 2.2.3 a conjugate base distribution makes it possible to analytically integrate out $\boldsymbol{\theta}$ and to sample only the allocation vector $\mathbf{Z}$. The parameter α is assigned a $Ga(4, 2)$ distribution, see Subsection 5.2.4 for motivation. The hyperparameters a, b and v are set to 1. An iteration of the MCMC algorithm consists of one conjugate Gibbs scan and three split-merge proposals as described by Dahl (2007). We found that the latter are beneficial in reducing the autocorrelation of the chain. More details on the MCMC sampler are given in Appendix C. When starting the sampler from a clustering with each observation in its own cluster, trace plots of the number of clusters indicate a quick convergence. After discarding the first 1000 iterations the algorithm is run for 50,000 iterations of which every 100th is used for the estimation of the posterior similarity matrix. The model is implemented in `R` where `C` functions are called for the time-demanding Gibbs sampling and split-merge steps. It takes about 8 minutes to run the model for one data set.

Comparison of optimization methods

First we take a look at the performance of the different optimization procedures mentioned in Subsection 5.4.3. Table 5.2 shows the average minimal value found for the posterior expectation of Binder's loss with different approaches. Using the `R` package `lpSolve` as done by Lau and Green (2007) to implement their algorithm it was not possible to apply the algorithm to all 400 observations, as the optimization problems required at each iteration got too large to be handled by the software. We therefore tested the approach

Table 5.2: Mean minimal value of posterior expectation of Binder's loss found with different optimization methods for the equal cluster size data.

Half of observations (n=200)					
	Draws	Comp	Avg	Avg&Grdy	Lau&Green
δ=0.5	5368	5378	5366	5365	5363
δ=1.0	6112	6095	6075	6070	6062
δ=1.5	2487	2175	2143	2142	2141
δ=2.0	1016	913	895	895	894

All observations (n=400)				
	Draws	Comp	Avg	Avg&Grdy
δ=0.5	21527	21561	21544	21544
δ=1.0	24477	24457	24315	24295
δ=1.5	10087	8842	8639	8629
δ=2.0	4465	3843	3739	3735

Draws refers to minimization over the MCMC sample, Comp and Avg to the minimization over all levels of the hierarchical clustering with complete/average linkage, Avg&Grdy to additional greedy optimization steps and Lau&Green to the optimization method proposed by Lau and Green (2007), which could not be applied to all observations.

by only considering a part of the posterior similarity matrix corresponding to the first half of observations from each true cluster. For these 200 observations it took the algorithm of Lau and Green about 30-40 minutes to finish for each data set. The algorithm did succeed in finding clusterings with the lowest value, but is closely followed by the hierarchical clustering with average linkage, which took less than a second to compute. The additional greedy assignment of single observations to other clusters makes the difference even smaller and takes only a few extra seconds. Note also that minimization of the criterion over the drawn clusterings does lead to minimal values that are much higher than for the other methods. Similar results have also been found for the unequal cluster size data and for the optimization of the criteria MAP and PEAR (results not shown). In the following we take the best clustering over all optimization methods as $\widehat{\mathbf{Z}}$ for each criterion.

Results

From the MCMC output of the Dirichlet process mixture model estimates $\widehat{\mathbf{Z}}$ are obtained by the following methods: Minimization of the posterior expectation of Binder's loss (Eqn. (5.20)) which is referred to as MinBinder, maximization of PEAR (Eqn. (5.24)) abbreviated as MPEAR, the complete linkage method for unknown K of Medvedovic et al. (2004) (MedvC) and by maximizing the posterior density (MAP). The relabel algorithm is applied using the approach proposed by Tadesse et al. (2006), i.e., discarding all draws $\mathbf{Z}^{(m)}$ where $K^{(m)}$ is not equal to the most often occurring number of groups. A finite mixture model is also fitted via the EM algorithm using the MCLUST procedure of Fraley and Raftery (2007) which is applied as implemented in the `R` package `mclust` with default settings. MCLUST chooses K via the BIC (see Subsection 2.1.6).

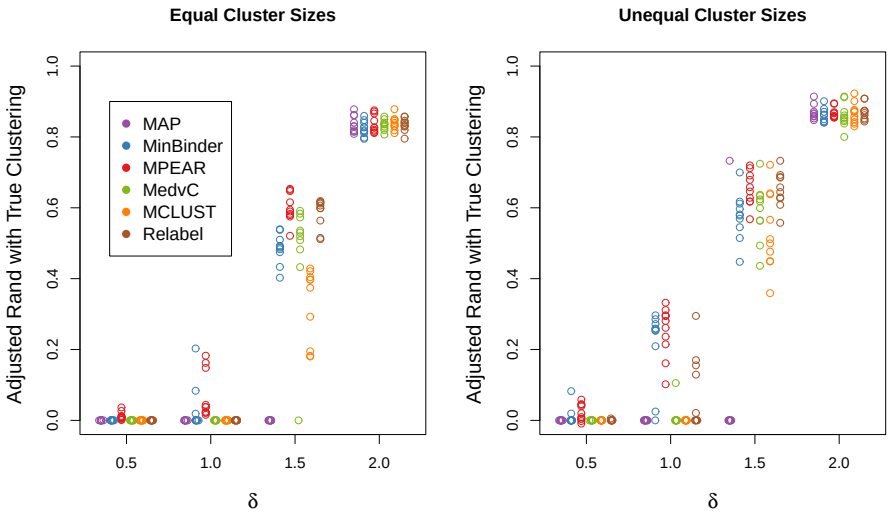

Figure 5.6: Adjusted Rand Index with true clustering for clusterings estimated with six different methods. Left: Data with clusters of equal sizes. Right: Data with clusters of unequal sizes.

To evaluate how close the $\widehat{\mathbf{Z}}$ of the different criteria are to the true clustering adjusted Rand indices of estimated and true clustering are computed. Figure 5.6 shows these for the data with equal and unequal cluster sizes. In the case of $\delta = 2$, i.e., for well separated clusters, all methods perform comparably. When the clusters are more overlapping MPEAR can be seen to give estimates closer to the true clustering than the other approaches. It gives notably better results than MinBinder, which it is meant to improve. This might be the case because the mentioned shrinkage effect comes into play. It is probably most beneficial if there are many π_{ij}'s close to 0.5, which is not the case for $\delta = 2$. For $\delta = 0.5$ most of the MCMC clusterings consist of only one cluster so that none of the estimation criteria can find meaningful cluster structure. The other methods also put all observations into one cluster for higher values of δ (MedvC, MCLUST and Relabel for δ equal to 1.0, MAP already for $\delta = 1.5$). In the case of MAP this can be explained by the high prior probability that the Dirichlet process places on allocating all observations to one component as can be inferred from formula (5.2). The likelihood has to be quite high to counteract this. The Relabel approach discards about 75%-90% of the MCMC allocation vectors for each data set but still performs very well in finding a clustering close to the true one. Figure B.1 in Appendix B gives the VI-distance to the true clustering for the simulation study. For low values of δ the main difference to the results with the adjusted Rand is that the VI-distance does not penalize putting all observations into one cluster as much as the adjusted Rand, so that now MinBinder and MPEAR give the worst results. For the higher values of δ especially the relative performance of MCLUST and Relabel is improved compared to Figure 5.6.

For the data sets with only one true cluster all criteria except MPEAR correctly assign all observations to one cluster. As noted PEAR will never be

Table 5.3: Mean number of clusters found in the simulation study.

	Equal Cluster Sizes					
	MAP	MinBinder	MPEAR	MedvC	MCLUST	Relabel
δ=0.5	1.0	1.4	6.5	1.0	1.0	1.1
δ=1.0	1.0	5.2	3.4	1.0	1.0	1.1
δ=1.5	1.0	89.1	12.7	6.6	3.4	8.2
δ=2.0	8.1	23.7	12.2	8.0	8.0	8.1
	Unequal Cluster Sizes					
	MAP	MinBinder	MPEAR	MedvC	MCLUST	Relabel
δ=0.5	1.0	2.5	5.9	1.0	1.0	1.6
δ=1.0	1.0	35.1	4.8	1.1	1.1	2.8
δ=1.5	1.7	72.2	14.3	5.2	4.6	7.5
δ=2.0	8.1	23.6	13.4	7.5	8.2	8.2
	Extreme Case Data					
	MAP	MinBinder	MPEAR	MedvC	MCLUST	Relabel
One Cluster	1.0	1.0	7.8	1.0	1.0	1.0
Singletons	1.0	1.0	7.7	1.0	4.3	1.0

maximal at one cluster and thus places some observations into other clusters, with most observations still being in one large cluster. For the data where all observations come from distinct components the MCMC output of the DP mixture model consists again mostly of clusterings with only one cluster and the criteria based on its output accordingly put all observations into one cluster. While this might look like a flaw of the DP mixture model, it can be argued that all observations in one cluster or all in singleton clusterings are just different ways of expressing that no cluster structures have been found in the data.

The mean number of clusters found by the different methods are shown in Table 5.3. It can be seen that for the higher values of δ the estimate obtained with MinBinder has a lot more clusters than the 8 truly present. This is also the case for MPEAR, but far less extreme. When the other methods do not

Table 5.4: Mean number of singletons and large clusters (more than 10 observations) for equal cluster size data.

	MPEAR		MinBinder	
	Singletons	Large Clusters	Singletons	Large Clusters
δ=0.5	2.5	2.2	0.1	1.0
δ=1.0	0.2	3.0	1.6	2.2
δ=1.5	2.0	8.1	66.5	8.9
δ=2.0	3.8	8.0	12.1	8.0

put all observations in one cluster their estimates have less than the true number of clusters for $\delta = 1.5$ and are on average approximately correct for $\delta = 2$.

The tendency of MinBinder and MPEAR to overestimate the number of clusters can be explained by the fact that many observations are put in singleton or in very small clusters. Table 5.4 shows the mean number of singletons and larger clusters for the two criteria. It can be seen that on average both methods are close to the correct number of (large) clusters for δ equal to 1.5 and 2. MinBinder produces many singletons. An example of this is given in Figure B.2 in the appendix. It shows the π_{ij} for an observation i that is put in a singleton cluster by both MinBinder and MPEAR and for an observation that is clustered by itself only by MinBinder. In the latter case it can be seen that the shrunken threshold t on π_{ij} of MPEAR leads to the observation being assigned to the correct cluster. In the simulation study comparing similarity measures for clusterings of Milligan and Cooper (1986) it was found that the Rand index will typically lead one to choose many clusters. Since MinBinder maximizes the posterior expected Rand index with the true clustering the results above are not surprising.

A sensitivity analysis of the simulation study concerning the number of MCMC iterations and different prior settings of the DP mixture model is

given in Appendix D. The results of the sensitivity analysis indicate that especially MPEAR is fairly robust to changes in the underlying model, at least if there is clear cluster structure in the data, i.e., for the cases $\delta = 1.5$ and $\delta = 2$. If there is no clear cluster structure the results are sensitive to the prior distribution of α.

5.5.2 Leukemia Data

Golub et al. (1999) obtained microarray gene expression measurements of bone marrow samples from leukemia patients. The patients have different subtypes of leukemia, B-cell acute lymphoblastic leukemia (ALL-B), T-cell acute lymphoblastic leukemia (ALL-T) and acute myeloid leukemia (AML), that differ in prognosis and treatment regime that should be applied. We investigate whether a Bayesian mixture model can recover the known group structure. The data are available in the *Bioconductor* repository (*www.bioconductor.org*, Gentleman et al., 2004) as package `golubEsets`. Considering only the training data set of Golub et al. (1999) the expression of 7129 genes has been measured on 38 patients. Some preprocessing steps as in Dudoit et al. (2002) are applied: setting measurements below 100 and above 16.000 to 100 and 16.000, respectively, and a logarithmic transformation. To reduce the dimensionality of the data we then compute principal components for the 1% genes with the largest variances across patients. The reason for using only highly variable genes is that these are the most likely to have a high between-group variation. And, as mentioned by McLachlan and Peel (2000, p. 239), if the groups are relatively well separated and the between-group variation dominates the within-group variation, the group structure should be represented by the projections on the first few principal axes.

As the first two principal components explain far more variation than the

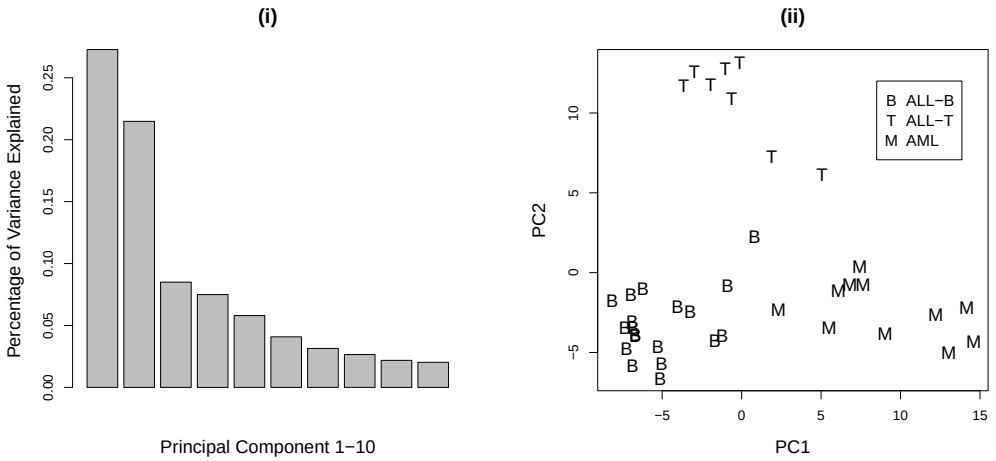

Figure 5.7: Leukemia data. (i) Percentage of total variance explained by first few principal components. (ii) Scatterplot of first two principal components.

others (see Figure 5.7 (i)) they will be used for the cluster analysis. Figure 5.7 (ii) plots the first two principal components. A group structure can be seen, even if the given annotation of the true subtypes is ignored.

As the clusters for the leukemia data are not necessarily spherical the simple DP mixture model used for the simulation study is replaced by a model with unrestricted covariance matrices. The model is

$$\begin{aligned} y_i|\mu_i,\Sigma_i &\sim N(\mu_i,\Sigma_i) \\ \mu_i,\Sigma_i|G &\sim G \\ G &\sim DP(\alpha, p(\mu,\Sigma)) \ , \end{aligned} \tag{5.29}$$

with base measure

$$\begin{aligned} p(\mu,\Sigma) &= p(\mu|\Sigma)p(\Sigma) \\ &= N(0, v^{-1}\Sigma)IW(c_0,C_0) \ . \end{aligned}$$

$IW(c_0,C_0)$ denotes the inverted Wishart distribution with expectation $E(\Sigma) = C_0/(c_0 - (d+1)/2)$ (which exists for $c_0 > (d+1)/2$). The hyperpa-

rameters are set to $v = 1$, $c_0 = 2.5$ and C_0 to a diagonal matrix containing the sample variances $s^2_{y_j}$, following Bensmail et al. (1997). As in the simulation study, α is given a $Ga(4, 2)$ prior. The model is fit using the function `DPdensity` from the `R` package `DPpackage` (Jara et al., 2009) with the number of MCMC iterations, burn-in and thinning as in the simulation study.

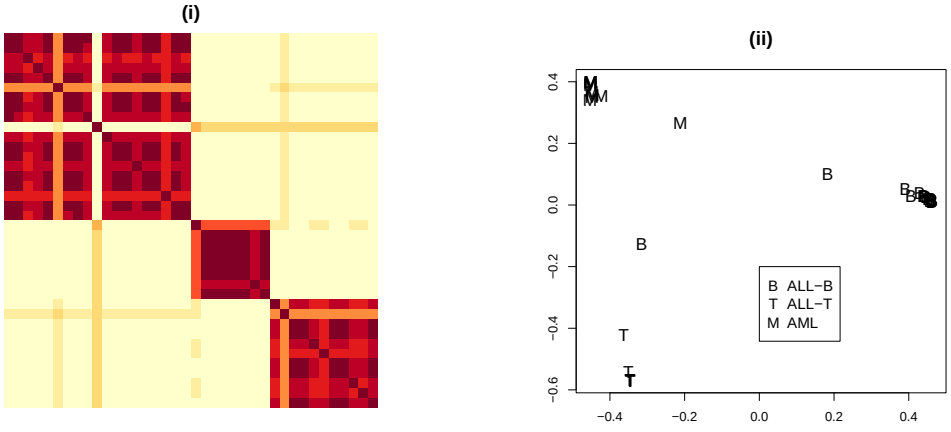

Figure 5.8: Leukemia data. (i) Heatmap of Posterior Similarity Matrix. (ii) Multidimensional scaling applied to Posterior Similarity Matrix.

Figure 5.8 shows the visualizations of the posterior similarity matrix computed from the MCMC sample of allocations $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$. The true subtypes have been well separated except for three observations. Figure 5.9 shows the clusterings estimated with the different methods. The MAP approach is able to perfectly recover the true subtypes, although from the heatmap in Figure 5.8 it can be seen that one of the ALL-B observations (this is the observation closest to the middle of Figure 5.7 (ii)) is only rarely in a cluster with the other ALL-B's. The second best solution in terms of adjusted Rand index and VI-distance is given by MPEAR, where this ALL-B observation is assigned its own cluster. The Relabel approach wrongly assigns the observation into a cluster with the ALL-T's. As before MinBinder has a tendency

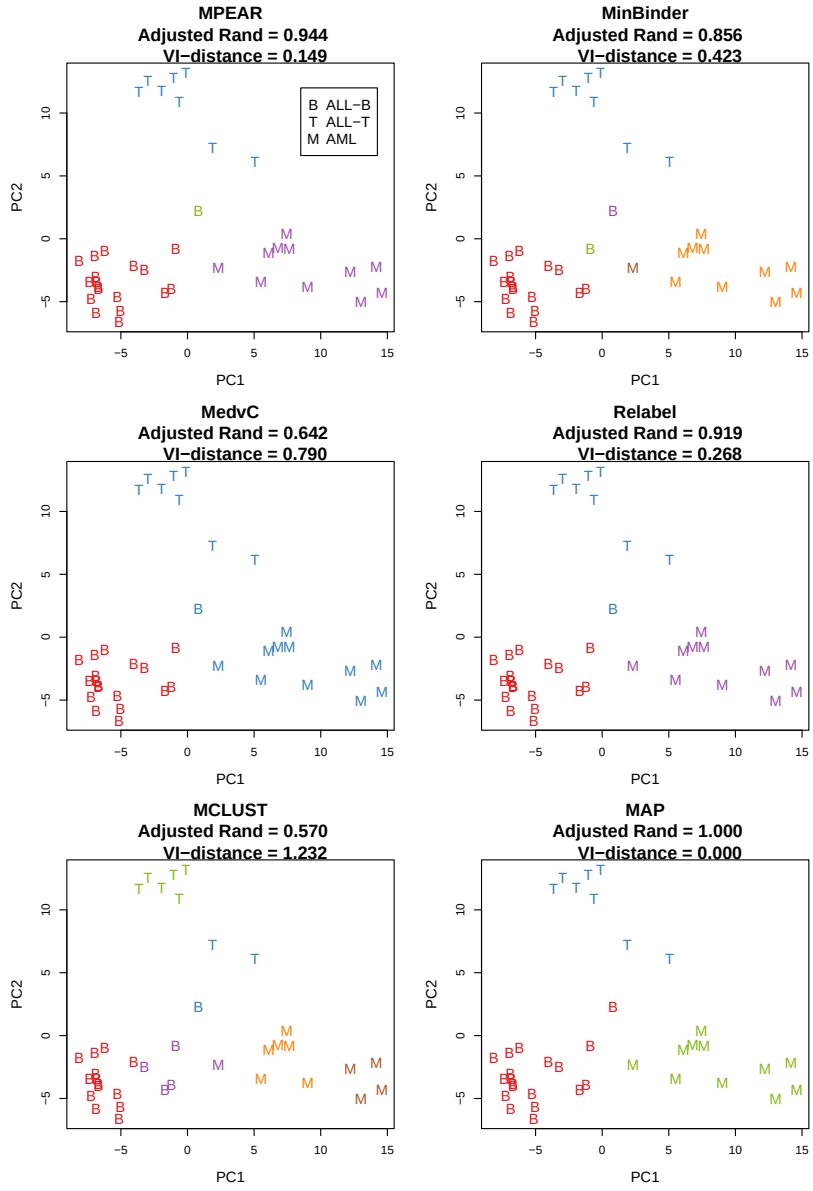

Figure 5.9: Clusterings of leukemia data. Letters denote true subtype, color indicates an estimated cluster. Adjusted Rand and VI-distance give similarity to true grouping of subtypes.

to produce singletons and assigns all three observations with unclear cluster status to their own cluster. For the leukemia data MedvC and MCLUST do not work very well and detect either too few or too many clusters. For

MedvC this can be explained by the fact that the one ALL-B observation and the AML and ALL-T observations are together in a cluster in about 2% of the MCMC clusterings. Hierarchical clustering with complete linkage and distances $1 - \pi_{ij}$ and cutting the dendrogram at a value of 0.99, as done by MedvC, then leads to these observations being in one cluster. MCLUST uses the BIC to decide between several possible variance structures and here chooses a structure that is too simple. With only 38 observations a asymptotic criterion like BIC is probably not justified.

Although the MAP clustering recovers the true subtypes it might be argued whether the clustering given by MPEAR is not preferable since it seems a sensible idea to assign the one ALL-B observation to a "status unclear" cluster.

5.5.3 Galactose Data

A Bayesian mixture model is applied to another gene expression data set, where it is the aim to cluster genes to obtain groups that are functionally related. The expression data are from a study on the galactose pathway (Ideker et al., 2001). Microarrays were used to measure mRNA concentrations under 20 different conditions in growing yeast, with the experiment being replicated four times. We use the same subset of 205 genes already employed by Medvedovic et al. (2004) and Qin (2006). The subset is chosen to reflect four functional categories of the Gene Ontology (Gene Ontology Consortium, 2000), which will be assumed to represent the true clustering. Figure 5.10 shows the mean expressions of the genes over the 20 conditions. There are two large groups containing 83 and 93 genes and two small ones with 14 and 15 genes. While Qin (2006) and Medvedovic et al. (2004) used the original data for clustering we will again use the first two principal com-

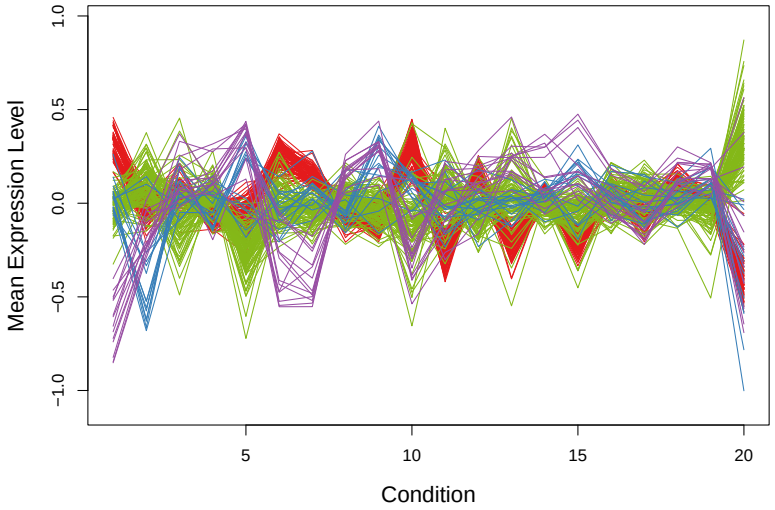

Figure 5.10: Mean expression levels of genes involved in the galactose pathway over different conditions. Colors indicate different functional categories of the Gene Ontology.

ponents as these explain a large part of the variation of the data and show clear cluster structures when plotted (see Figure B.3 in the appendix). The Dirichlet process mixture model (5.29) used for the leukemia data is fitted with the same prior settings and number of MCMC iterations to the PCs derived from each single experimental replicate and from the average expression over the four replicates.

Table 5.5 gives the similarity measures with the true clustering for the clusterings obtained with the different estimation approaches. Overall we obtain better results using principal components than reported by Medvedovic et al. and Qin for the original data. For the mean expression data MinBinder and MPEAR give the best solution, although all methods perform relatively well. For the single replicates MinBinder and MPEAR do not perform very well. The reason is that here both put many observations into singletons,

Table 5.5: Similarity measure with true clustering for yeast galactose data. (Results for replicates are averages over the four replications.)

	Adjusted Rand index					
	MAP	MinBinder	MPEAR	MedvC	MCLUST	Relab
Mean Expr.	0.965	0.965	0.965	0.960	0.937	0.965
Replicates	0.922	0.850	0.878	0.913	0.841	0.866
	VI-distance					
	MAP	MinBinder	MPEAR	MedvC	MCLUST	Relab
Mean Expr.	0.248	0.224	0.224	0.241	0.379	0.224
Replicates	0.439	0.678	0.602	0.489	0.633	0.586

MinBinder again some more than MPEAR. Similar to the leukemia data for the single replicates, MAP gives clusterings closest to the true one.

With the original data, instead of using an average expression vector, the four replicates could be modeled with an additional hierarchical level in model (5.29). A mean expression vector m_i then replaces y_i in (5.29) and the jth repetition of the ith gene expression vector y_{ij} is, for example, modeled by

$$y_{ij}|m_i \sim N(m_i, \psi_i I) \ , \tag{5.30}$$

thus allowing to take gene-specific variances ψ_i into account. Medvedovic et al. (2004) consider such a model and demonstrate its usefulness in the presence of heteroscedastic genes in a simulation study. For the galactose data they find that not much is gained with the inclusion of a hierarchical level of the form (5.30), so that the genes seem to have similar variances. The use of average expression profiles thus seems to be justified here.

5.5.4 Iris Data

The iris data from Anderson (1935), first analyzed by Fisher (1936), are a well known data set in multivariate analysis. The data consist of four

measurements (length and width of the sepals and length and width of the petals) on 150 irises, with 50 observations from each of the species *iris setosa*, *versicolor*, and *virginica*. Fitting either the DP mixture model (5.29) with the previous prior settings or MCLUST does lead to a clustering where *versicolor* and *virginica* are merged into one cluster, no matter which of the estimation methods is applied. Extending the model (5.29) with hyperpriors on v and C_0 leads to better results. The priors used are $v \sim Ga(1,1)$ and $C_0 \sim W(g_0, G_0)$ with $W(g_0, G_0)$ being a Wishart distribution with expectation G_0/g_0. Along the lines of Stephens (1997) the remaining hyperparameters are set to $c_0 = d + 1 = 5$, $g_0 = c_0/10$ and G_0 to a diagonal matrix with entries $10/R_j^2$, R_j^2 being the range of the jth $\mathbf{y}$ variable. The extended model can be fit with the `DPdensity` function of the `DPpackage` package.

Table 5.6: Contingency tables of iris grouping with clusterings estimated by different methods.

	MedvC, MAP			MinBinder, MPEAR, Relabel		
Cluster	*setosa*	*versicolor*	*virginica*	*setosa*	*versicolor*	*virginica*
1	50	0	0	50	0	0
2	0	46	1	0	45	0
3	0	4	49	0	5	41
4	-	-	-	0	0	9

The posterior similarity matrix that is obtained from the MCMC output of the extended model has already been shown in Figure 5.5. It can be seen that the *setosa* group is well separated from the other two, that there is some overlap between *versicolor* and *virginica* and that there are two subgroups of *virginica*. Applying the estimation methods to the PSM yields two clusterings. Their contingency tables with the true iris grouping are shown in Table 5.6. The criteria MAP and MedvC find a clustering with three clusters with

some observations interchanged between the *versicolor* and *virginica* group. In the clustering with four clusters found by MinBinder, MPEAR and Relabel the smaller subgroup of the *virginica* cluster is made up of the observations 6, 8, 18, 19, 23, 26, 30, 31, 32. The original reason of Anderson (1935) for collecting the iris data was to look for signs of continuing evolution. Looking at the scatterplots in Figure 5.11 it seems plausible that the smaller cluster corresponds to a potential subspecies of *iris virginica*, as the observations in it have large values for petal width, petal length and sepal length and there is a gap between them and the other *virginica* observations. The distributions of sepal width is similar between the two clusters.

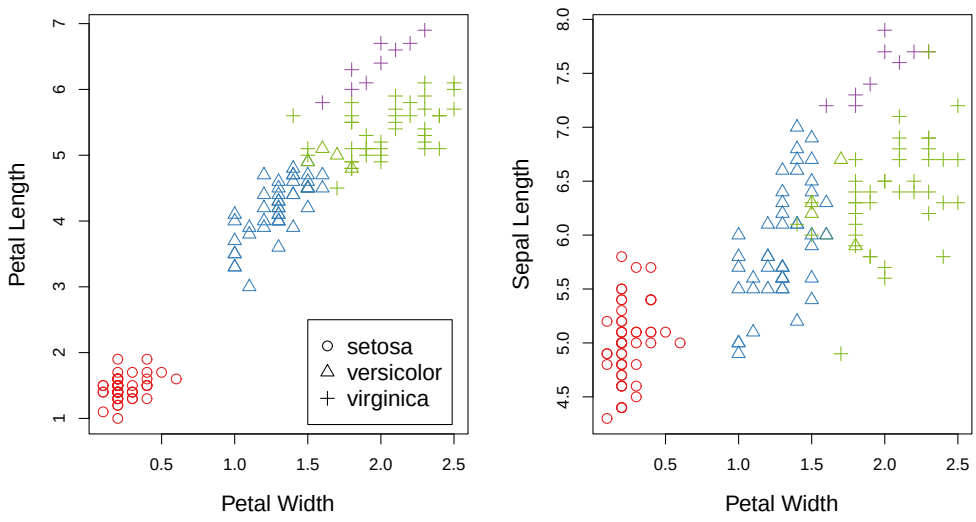

Figure 5.11: Iris data. Petal width against petal length and sepal length. Color indicates estimated clusters.

The subgroups of *virginica* have been identified previously. They are mentioned by Wilson (1982) who used an exploratory data analysis method based on sound, where p-dimensional observations are transformed into a melody of p notes. McLachlan and Peel (2000, Chapter 3.11) use the EM algorithm

to fit a mixture of two normals with unrestricted covariance matrices to only the data of the *virginica* group and find that the splitting of the data into the discussed two subgroups corresponds to a local mode of the likelihood that is probably not spurious. They find, however, that a formal test of $H_0 : K = 1$ vs. $H_1 : K = 2$ is not significant. In contrast, the model fitted by us strongly suggests the existence of the two *virginica* clusters, as they are present in most of the MCMC samples. This can be seen clearly in Figure 5.5. The difference is possibly due to the use of the hyperprior on C_0, i.e., $C_0 \sim W(g_0, G_0)$. Use of a hyperprior is equivalent to assuming that the clusters have different but similar covariance matrices. From Figure 5.11 and the other scatterplots of the variables (not shown) the assumption of similar covariance matrices seems reasonable. As the use of a hyperprior on C_0 seems plausible and the model incorporating such a prior strongly suggest four clusters there is some support for the existence of two subspecies of *virginica*.

Chapter 6

Conclusions and Outlook

In this thesis Bayesian mixtures, especially finite mixtures and Dirichlet process mixtures, have been investigated. One possible application of these models is as a semiparametric extension of standard parametric models. The applicability of Bayesian mixtures for such a flexible modeling of distributions has been demonstrated both theoretically and practically. The former was done by stating consistency results and the latter by modeling the random effects distribution in a logistic regression model. In this application, which considered the goalkeeper's effect in saving a penalty, the Dirichlet process mixture model led to results very similar to the ones resulting from a normal model. The DP mixture model nevertheless provided a useful sensitivity analysis of the normal model and indicated that the tails of the random effects distribution should be somewhat more pronounced. From an application point of view the used hierarchical model, with its associated shrinkage effect, is more appropriate than using just the number or percentage of saved penalties for estimating the ability of a goalkeeper.

When using a Bayesian mixture for cluster analysis, the arbitrariness of cluster labels leads to the label-switching problem in the MCMC fitting of the model. The problem gets severe when the number of clusters is allowed

to vary. Two new approaches to this problem have been proposed in this thesis. The first consists of a more generally applicable variant of the relabeling algorithm of Stephens (2000). The variant is more general, as it applies to drawn clusterings and not drawn parameter values. Therefore it does not depend on the specific form of the component distributions. The second approach is based on pairwise posterior probabilities and is an improvement of a commonly used loss function due to Binder (1978). Minimization of this loss has been shown to be equivalent to maximizing the posterior expected Rand index with the true clustering. As the adjusted Rand index is preferable to the raw index, the maximization of the posterior expected adjusted Rand has been proposed. The resulting PEAR criterion could be shown to possess a shrinkage property and performed well in a simulation study and in an application to two gene expression data sets, where estimated clusterings closer to the truth than the ones resulting from minimizing Binder's loss could be found. The number of clusters was not assumed to be known for these data sets. Although the relabeling algorithm was designed for a fixed number of clusters, it has been applied after conditioning on a MAP estimate of the number of clusters. Despite the fact that the conditioning required to discard 75% - 90% of the MCMC clusterings, the relabeling approach worked almost as good as PEAR. Both approaches compared favorably to the clusterings obtained using an ad hoc criterion of Medvedovic et al. (2004) and MCLUST. The MAP clustering did not give good results for overlapping clusters in the simulation study, but performed well for the gene expression data. Similarly to Stephens' relabeling algorithm, a specific formula has to be used for the MAP in each mixture model, whereas the other methods can be applied without modification to the MCMC output of a Bayesian mixture model. We would therefore generally recommend to use PEAR or the

relabeling algorithm instead of Binder's loss or Medvedovic's approach. A model-specific MAP also seems to work well if clusters are not overlapping.

For Fishers iris data the new postprocessing methods, together with the assumption of similar covariances in the clusters, lead to the (re-)discovery of two subgroups of *iris virginica*, which might be an indication of continuing evolution in these plants.

In the optimization of Binder's loss and PEAR it turned out that hierarchical clustering with $1 - \pi_{ij}$ as distance and average linkage is a quick way to get a good approximation to the optimum. The attained solution can be further improved with some greedy optimization steps. For minimization of the posterior expectation of Binder's loss this approach gave estimated clusterings with almost as low values as obtained with the optimization algorithm of Lau and Green (2007), but took only a few seconds to compute, instead of 30 minutes. Although frequently proposed in the literature, optimizing the criteria only over the MCMC sample of clusterings performed worse and seems not to be a very good way to obtain an estimate $\widehat{\mathbf{Z}}$.

An issue not yet discussed is the scalability of the proposed methods. Regarding the dimension p of each single observation the clustering models in Section 5.5 could be applied without problems when p is in the order of a few tens. In that case a model with a restricted covariance structure, like (5.28), should be preferred to a model with an unrestricted covariance matrix, like (5.29). As demonstrated with the two gene expression data sets using principal components to reduce p can work very well for clustering. Although this does not have to be the case, dimension reduction with principal components is a simple approach that usually seems worth trying. Regarding the number of observations n, Medvedovic et al. (2004) and Dahl (2006) apply clustering methods based on Dirichlet process mixtures and the pos-

terior similarity matrix to data sets with $n \approx 10,000$. Besides running times of several hours to few days on a desktop computer they do not report any problems. As discussed in Subsection 5.4.2, for large n it might be preferable to avoid computation of the posterior similarity matrix and compute PEAR using (5.23) rather than (5.24). The proposed relabeling algorithm is computationally less demanding and thus more suitable for large n than Stephens' approach. The most important limiting factor for the algorithm is probably the number of clusters K, as each iteration requires solving M assignment problems with running times of the order $\mathcal{O}(K^3)$.

The work on PEAR in this thesis could be extended by considering the minimization of the posterior expectation of the distance between estimated and true clustering, using distance measures other than the adjusted Rand, e.g., the discussed "variation of information"-distance. An empirical comparison of the proposed relabeling algorithm with Stephens' original approach might also be interesting.

Bayesian mixtures are a very active area of research. The main focus of this thesis has been on cluster analysis and the modeling of a single distribution. The interesting modeling of several related distributions has only been mentioned briefly in Subsection 2.2.4. Much work is also done concerning Bayesian mixtures of regression or time series models, see for example Frühwirth-Schnatter (2006, Chapters 8 and 10-13) and Bornkamp et al. (2010). A large number of both theoretical and applied articles have also been published in recent years in the general area of Bayesian nonparametrics, to which methods based on the Dirichlet process are counted. Developments include extensions of the Dirichlet process, alternatives, for example, generalized gamma process priors (Lijoi et al., 2007), new computational approaches and asymptotic theory. An overview of recent results can be found in Hjort

et al. (2010). Many interesting statistical methods are still to be expected from this line of research.

Appendix A

Additional Proof

Proof of Lemma 5.3:

The conditions on $\pi_{(\ell^*)}$ for $PEAR^*$ of equation (5.25) to take its maximum are considered. This is done by considering under what conditions $PEAR^*(\ell)$ is greater than $PEAR^*(\ell-1)$. Recall that $\pi_{(i)}$ is the ith largest π_{ij}. Then

$$\begin{aligned}
PEAR^*(\ell) &\geq PEAR^*(\ell-1) \\
\iff \frac{\sum_{i=1}^{\ell}\pi_{(i)} - \ell\,\bar{\pi}_{..}}{\frac{1}{2}\ell\,(1-2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} &\geq \frac{\sum_{i=1}^{\ell-1}\pi_{(i)} - (\ell-1)\,\bar{\pi}_{..}}{\frac{1}{2}(\ell-1)\,(1-2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} \\
\iff \pi_{(\ell)}\left[(\ell-1)(1-2\bar{\pi}_{..}) + \binom{n}{2}\bar{\pi}_{..}\right] &\geq (1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}\pi_{(i)} + \binom{n}{2}\bar{\pi}_{..}^2 \quad \text{(A.1)} \\
\iff \pi_{(\ell)}\binom{n}{2}\bar{\pi}_{..} - \binom{n}{2}\bar{\pi}_{..}^2 &\geq (1-2\bar{\pi}_{..})\left[\sum_{i=1}^{\ell-1}\pi_{(i)} - (\ell-1)\pi_{(\ell)}\right] \\
\iff \binom{n}{2}\bar{\pi}_{..}(\pi_{(\ell)} - \bar{\pi}_{..}) &\geq (1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}(\pi_{(i)} - \pi_{(\ell)}) \quad \text{(A.2)}
\end{aligned}$$

Obtaining (A.1) from the previous line requires multiplication with the denominators, expansion and deletion of terms occurring on both sides of the equation. For $\bar{\pi}_{..} \leq 0.5$ (A.2) is decreasing on the left hand side with rising ℓ and increasing on the right hand side, so that $PEAR^*$ has a unique maximum (or two maxima at adjacent values). The uniqueness of the maximum for $\bar{\pi}_{..} \leq 0.5$ is a bit more complicated and is shown later.

The threshold t can be determined by setting (A.2) to equality

$$\binom{n}{2}\bar{\pi}_{..}(t-\bar{\pi}_{..}) \;=\; (1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}(\pi_{(i)}-t)\ . \tag{A.3}$$

Since $\sum_{i=1}^{\ell-1}(\pi_{(i)}-t)>0$ it follows that

$$\begin{aligned} t &> \bar{\pi}_{..} \qquad \text{if} \quad \bar{\pi}_{..}<0.5 \\ t &= \bar{\pi}_{..} \qquad \text{if} \quad \bar{\pi}_{..}=0.5 \\ t &< \bar{\pi}_{..} \qquad \text{if} \quad \bar{\pi}_{..}>0.5\ . \end{aligned}$$

To prove the relation of t to 0.5 we solve (A.3) for t and show that for $\bar{\pi}_{..}<0.5$ it is smaller than 0.5:

$$\begin{aligned} \frac{(1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}\pi_{(i)}+\binom{n}{2}\bar{\pi}_{..}^2}{(1-2\bar{\pi}_{..})(\ell-1)+\binom{n}{2}\bar{\pi}_{..}} &< \frac{1}{2} \\ \iff 2(1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}\pi_{(i)}+2\binom{n}{2}\bar{\pi}_{..}^2 &< (1-2\bar{\pi}_{..})(\ell-1)+\binom{n}{2}\bar{\pi}_{..} \\ \iff (1-2\bar{\pi}_{..})\left[\sum_{i=1}^{\ell-1}2\pi_{(i)}-(\ell-1)\right] &< \binom{n}{2}\bar{\pi}_{..}-2\binom{n}{2}\bar{\pi}_{..}^2 \\ \iff (1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}(2\pi_{(i)}-1) &< (1-2\bar{\pi}_{..})\binom{n}{2}\bar{\pi}_{..} \\ \iff \sum_{i=1}^{\ell-1}(2\pi_{(i)}-1) &< \binom{n}{2}\bar{\pi}_{..}\ . \end{aligned}$$

The maximum that the term on the left can take is $(\ell-1)$, this is the case if the first $(\ell-1)$ $\pi_{(i)}$ are equal to 1. As $\binom{n}{2}\bar{\pi}_{..}=\sum_i\pi_{(i)}$, the term on the right is at least as large. Equality holds, if $\pi_{(i)}=1$ for $i\leq\ell-1$ and all other $\pi_{(i)}=0$.

For $\bar{\pi}_{..}>0.5$ the direction of the inequality is changed when dividing by $(1-2\bar{\pi}_{..})$, so that t is larger than 0.5 in this case.

It remains to show that $PEAR^*$ has a unique maximum for $\bar{\pi}_{\cdot\cdot} > 0.5$. Rewriting (A.1) yields

$$\pi_{(\ell)} \geq \frac{(1-2\bar{\pi}_{\cdot\cdot})\sum_{i=1}^{\ell-1}\pi_{(i)} + \binom{n}{2}\bar{\pi}_{\cdot\cdot}^2}{(1-2\bar{\pi}_{\cdot\cdot})(\ell-1) + \binom{n}{2}\bar{\pi}_{\cdot\cdot}} \quad . \tag{A.4}$$

The left hand side of (A.4) is decreasing with rising ℓ and the right hand side is equal to $\bar{\pi}_{\cdot\cdot}$ for $\ell = 1$ and $\ell = \binom{n}{2}$. To determine the behavior of the r.h.s. of (A.4) for the intermediate values of ℓ it is considered under what conditions it is decreasing. The equation

$$\frac{(1-2\bar{\pi}_{\cdot\cdot})\sum_{i=1}^{\ell}\pi_{(i)} + \binom{n}{2}\bar{\pi}_{\cdot\cdot}^2}{(1-2\bar{\pi}_{\cdot\cdot})\ell + \binom{n}{2}\bar{\pi}_{\cdot\cdot}} \leq \frac{(1-2\bar{\pi}_{\cdot\cdot})\sum_{i=1}^{\ell-1}\pi_{(i)} + \binom{n}{2}\bar{\pi}_{\cdot\cdot}^2}{(1-2\bar{\pi}_{\cdot\cdot})(\ell-1) + \binom{n}{2}\bar{\pi}_{\cdot\cdot}}$$

can be shown to be equivalent to (A.4). Then $PEAR^*(\ell) \geq PEAR^*(\ell-1)$ iff the r.h.s. of (A.4) is decreasing. Since the l.h.s. of (A.4) is in general decreasing the maximum has to be unique.

Appendix B

Additional Graphs and Tables

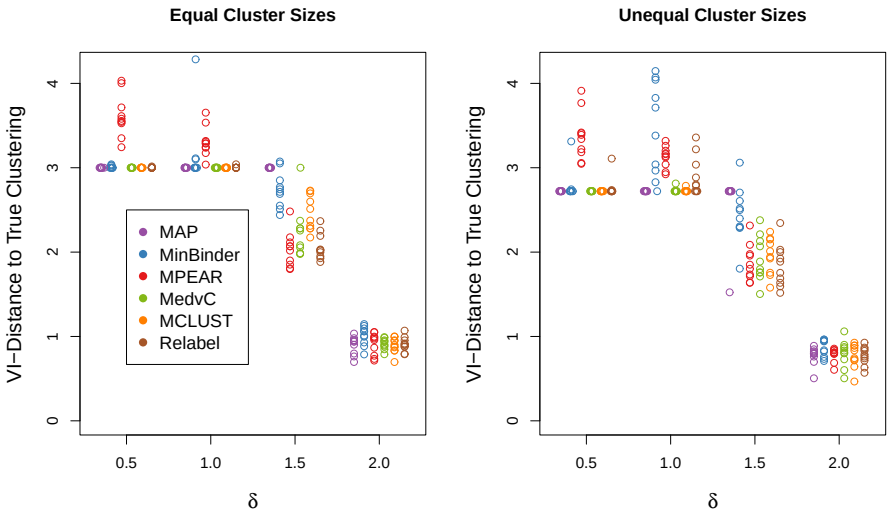

Figure B.1: VI-distance to true clustering for clusterings of simulation study. Left: Data with clusters of equal sizes. Right: Data with clusters of unequal sizes.

Goalkeeper	Rank	$P(\text{Saving}\|y)$ with 95% CI	% Saved	# Saved	# Penalties
Kargus, Rudolf	1	0.260 [0.189, 0.354]	0.329	23	70
Enke, Robert	2	0.243 [0.166, 0.357]	0.407	11	27
Pfaff, Jean-Marie	3	0.249 [0.160, 0.426]	0.545	6	11
Köpke, Andreas	4	0.234 [0.165, 0.329]	0.317	13	41
Radenkovic, Petar	5	0.233 [0.161, 0.333]	0.353	12	34
⋮	⋮	⋮	⋮	⋮	⋮
Melka, Michael	59	0.204 [0.126, 0.304]	1.000	1	1
⋮	⋮	⋮	⋮	⋮	⋮
Teupel, Gerhard	134	0.192 [0.113, 0.287]	0.000	0	1
⋮	⋮	⋮	⋮	⋮	⋮
Lehmann, Jens	221	0.186 [0.123, 0.259]	0.189	7	37
Kahn, Oliver	222	0.186 [0.126, 0.253]	0.172	10	58
⋮	⋮	⋮	⋮	⋮	⋮
Schmadtke, Jörg	284	0.167 [0.101, 0.234]	0.098	4	42
Rynio, Jürgen	285	0.164 [0.089, 0.235]	0.074	2	27
Müller, Manfred	286	0.163 [0.085, 0.237]	0.040	1	25
Junghans, Walter	287	0.163 [0.086, 0.236]	0.042	1	24
Maier, Sepp	288	0.162 [0.105, 0.221]	0.130	9	69

Table B.1: Ranking of goalkeepers based on the Dirichlet process mixture model (3.3). The ranking is determined by the average rank during the MCMC run and not by the average posterior saving probability.

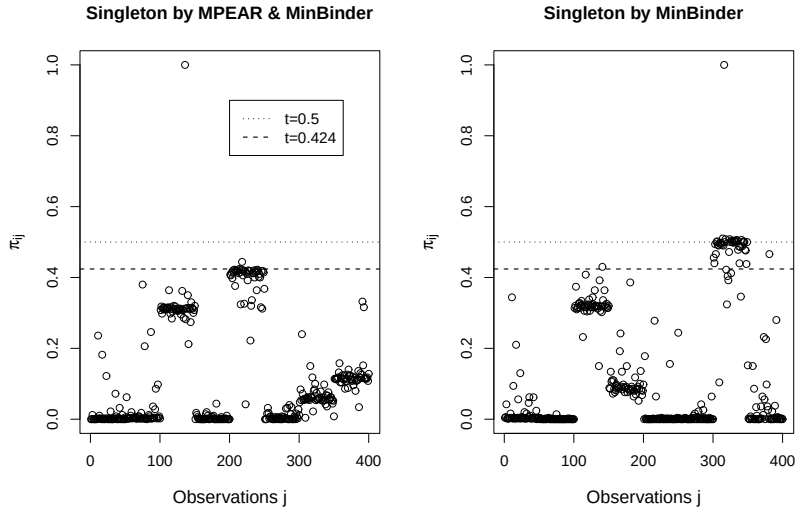

Figure B.2: Pairwise posterior probabilities π_{ij} for two observations i. Equal cluster size data with $\delta = 2$ and $\bar{\pi}_{\cdot\cdot}$=0.119. Left: Observation is put into its own cluster by MPEAR and MinBinder. Right: Observation is put into its own cluster only by MinBinder. Lines indicate thresholds t for MPEAR (- - -) and MinBinder ($\cdots$), t is described in Subsection 5.4.2.

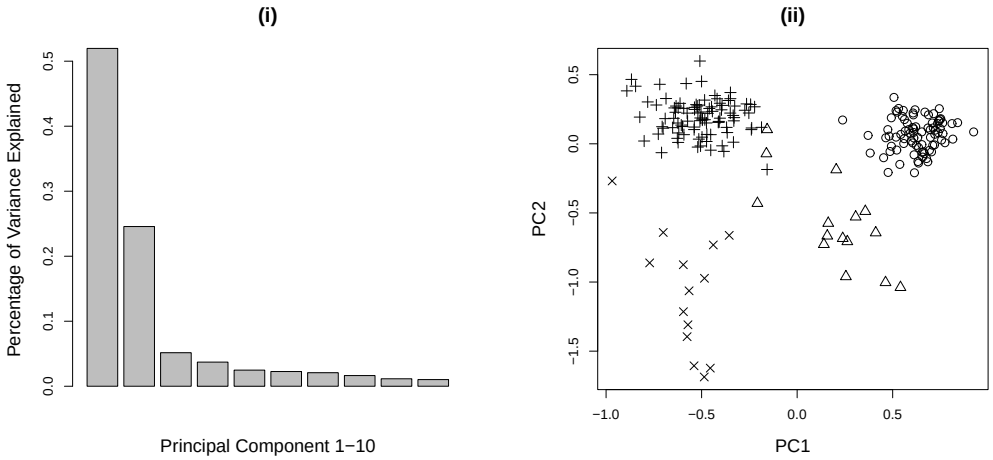

Figure B.3: Principal components of mean expression galactose data. (i) Percentage of total variance explained by first few principal components. (ii) Scatterplot of first two principal components. Plot characters denote the four Gene Ontology groups.

Appendix C

Details on MCMC Sampler

This section describes the MCMC sampler used for fitting the simple Dirichlet process mixture model (5.28) in Subsection 5.5.1.

The conditional distribution of Z_i given all other indicators $\mathbf{Z}_{-i}$ and $\mathbf{y}$ is given by

$$P(Z_i = k|\mathbf{Z}_{-i}, \mathbf{y}) \propto \frac{n_{k,-i}}{\alpha+n-1} \int N(y_i|\mu, \sigma^2) p(\mu, \sigma^2|\mathbf{y}_{k,-i})\, d\mu\, d\sigma^2 \quad \text{(C.1)}$$

$$P(Z_i = K+1|\mathbf{Z}_{-i}, \mathbf{y}) \propto \frac{\alpha}{\alpha+n-1} \int N(y_i|\mu, \sigma^2) p(\mu, \sigma^2)\, d\mu\, d\sigma^2 \ , \quad \text{(C.2)}$$

where K is the number of clusters in $\mathbf{Z}_{-i}$, $n_{k,-i}$ is the number of $\mathbf{Z}_{-i} = k$ and $\mathbf{y}_{k,-i}$ are the corresponding observations. Conditioning on current draws of any of the hyperparameters α, b and v that is assigned a prior distribution is also assumed. Since $p(\mu|\sigma^2)p(\sigma^2) = N_p(0, \sigma^2 v^{-1} I_p) InvGa(a, b)$ the integral in (C.2) can be solved by first integrating with respect to μ, leading to $p(y_i|Z_i = K+1, \sigma^2) = N_p(0, \sigma^2(1+v^{-1})I_p)$ and then using results of Bernardo and Smith (1994, p.140) to obtain

$$p(y_i|Z_i = K+1) = t_p(0, \frac{b}{a}(1+v^{-1})I, 2a) \ ,$$

where $t_p(\eta, \Sigma, \nu)$ is the p-dimensional Student t-distribution with expectation η (for $\nu > 1$) and variance $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$).

The integral in (C.1) can be solved by first employing standard results in Bayesian inference to give $p(\mu|\sigma^2, \mathbf{y}_{k,-i})p(\sigma^2|\mathbf{y}_{k,-i}) = N_p(\mu^*, \sigma^2 v^{*-1} I) InvGa(a^*, b^*)$, where

$$\begin{aligned} \mu^* &= \frac{n_{k,-i}}{v + n_{k,-i}} \bar{\mathbf{y}}_{k,-i} \\ v^* &= v + n_{k,-i} \\ a^* &= a + d n_{k,-i}/2 \\ b^* &= b + \frac{1}{2}[\sum_{\substack{j \neq i \\ Z_j = k}} y_j^T y_j - v^* \mu^{*T} \mu^*] \ . \end{aligned}$$

Applying the same reasoning as above one then obtains

$$p(y_i|Z_i = k, \mathbf{Z}_{-i}, \mathbf{y}_{k,-i}) = t_p(\mu^*, \frac{b^*}{a^*}(1 + v^{*-1})I, 2a^*) \ .$$

If the hyperparameter α is assigned a $Ga(\delta_1, \delta_2)$ prior the Gibbs sampling scheme of Escobar and West (1995) is employed.

A $Beta(v_1, v_2)$ prior on $1/(1+\alpha)$ induces the Beta of the second kind prior on α given in (5.6) leading to a full conditional depending only on K and n

$$p(\alpha|K, n) \propto \frac{\alpha^{K+v_2-1}}{(1+\alpha)^{(v_1+v_2)}} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}$$

which is updated with a random walk Metropolis algorithm on $\log(\alpha)$.

If v is assigned a $Ga(\gamma_1, \gamma_2)$ prior its full conditional is given by a

$$Ga(\gamma_1 + \frac{K \cdot d}{2}, \gamma_2 + \frac{1}{2} \sum_{k=1}^{K} \sigma_k^{-2} \cdot \mu_k^T \mu_k) \ ,$$

where μ_k, σ_k^2 are sampled for each cluster from the $N_p(\mu^*, \sigma^2 v^{*-1} I) InvGa(a^*, b^*)$ distribution given above. Similarly, if b is assigned a $Ga(\eta_1, \eta_2)$ prior the full conditional is a $Ga(\eta_1 + K \cdot a, \eta_2 + \sum_{k=1}^{K} \sigma_k^{-2})$.

Appendix D

Sensitivity Analysis of Simulation Study

This section gives a sensitivity analysis of the simulation study in Subsection 5.5.1 concerning the number of MCMC iterations and different prior settings of the Dirichlet process mixture model (5.28).

To evaluate the effect of the number of MCMC iterations some of the data sets have been fitted with twice the number of iterations, which did not improve the results (data not shown). More iterations seemed only to be beneficial if the optimization of the criteria is done solely over the drawn allocations $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \ldots, \mathbf{Z}^{(M)}$, where in some case better scoring clusterings could be found.

To investigate the effect that the underlying clustering model has on the estimated clusterings the DP mixture model is also fit to the simulated data with different priors for the hyperparameters. The first modification consists of placing additional $Ga(1,1)$ priors on both b and v. For the second setting the $Ga(4,2)$ on α is replaced by a vague prior. As mentioned in Subsection 5.2.4 in a clustering context a good choice for a vague prior seems to be to let $P(Z_i = Z_j) = 1/(1+\alpha) \sim Beta(1,1)$. The last setting has a fixed $\alpha = 4$.

Table D.1 shows the average adjusted Rand indices that resulted from

Table D.1: Average adjusted Rand index with the true clustering for equal cluster size data and different prior settings.

			$\delta = 1.5$		
Prior	MAP	MinBinder	MPEAR	MedvC	Relabel
Standard	0.000	0.485	0.601	0.476	0.586
Gamma b,v	0.000	0.393	0.568	0.456	0.547
Vague α	0.000	0.414	0.547	0.326	0.423
Fixed α	0.000	0.428	0.592	0.550	0.593
			$\delta = 2$		
Prior	MAP	MinBinder	MPEAR	MedvC	Relabel
Standard	0.836	0.824	0.837	0.833	0.835
Gamma b,v	0.784	0.793	0.821	0.827	0.829
Vague α	0.838	0.826	0.840	0.835	0.838
Fixed α	0.831	0.820	0.831	0.836	0.837

applying the estimation methods to the output of the model with the different priors for the equal cluster size data and δ equal to 1.5 and 2. For $\delta = 2$ the results are relatively robust for all criteria. In the case of $\delta = 1.5$ the average adjusted Rand indices are generally a bit worse for the "Gamma b, v" and "Vague α" setting, indicating that here the added prior flexibility makes it harder to find the true cluster structure. The results of MPEAR are the least affected by the different priors and still lead to the best average result. The results for δ equal to 0.5 and 1 do not change much compared to the standard prior (data not shown), except that for $\delta = 1$ and the "Vague α" setting none of the criteria can find any cluster structure and that the results are better for MinBinder and MPEAR for the "Fixed α" setting.

A similar pattern is found for the unequal cluster size data while the results for the extreme case data sets are not affected by the different priors.

Appendix E

R Package `mcclust`

`mcclust-package` *Process MCMC Sample of Clusterings.*

Description

Implements methods for processing a sample of (hard) clusterings, e.g., the MCMC output of a Bayesian clustering model. Among them are methods that find a single best clustering to represent the sample, which are based on the posterior similarity matrix or a relabeling algorithm.

Details

Package:	mcclust
Type:	Package
Version:	1.0
Date:	2009-03-12
License:	GPL (>= 2)
LazyLoad:	yes

Most important functions:

`comp.psm` for computing posterior similarity matrix (PSM). Based on the PSM `maxpear` and `minbinder` provide several optimization methods to find a clustering with maximal posterior expected adjusted Rand index with the true clustering or one that minimizes the posterior expectation of a loss function by Binder (1978). `minbinder` provides the optimization algorithm of Lau and Green.

`relabel` contains the relabeling algorithm of Stephens (2000).

`arandi` and `vi.dist` compute distance functions for clusterings, the (adjusted) Rand index and the entropy-based variation of information distance.

Author(s)

Arno Fritsch

Maintainer: Arno Fritsch <arno.fritsch@tu-dortmund.de>

References

Binder, D.A. (1978) Bayesian cluster analysis, *Biometrika* **65**, 31–38.

Fritsch, A. and Ickstadt, K. (2009) An improved criterion for clustering based on the posterior similarity matrix, *Bayesian Analysis*, **4**, 367–392.

Lau, J.W. and Green, P.J. (2007) Bayesian model based clustering procedures, *Journal of Computational and Graphical Statistics* **16**, 526–558.

Stephens, M. (2000) Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society Series B*, **62**, 795–809.

Examples

```
data(cls.draw2)
# sample of 500 clusterings from a Bayesian cluster model
tru.class <- rep(1:8,each=50)
# the true grouping of the observations
psm2 <- comp.psm(cls.draw2)
# posterior similarity matrix

# optimize criteria based on PSM
mbind2 <- minbinder(psm2)
mpear2 <- maxpear(psm2)

# Relabeling
k <- apply(cls.draw2,1, function(cl) length(table(cl)))
max.k <- as.numeric(names(table(k))[which.max(table(k))])
relab2 <- relabel(cls.draw2[k==max.k,])

# compare clusterings found by different methods with
# true grouping
arandi(mpear2$cl, tru.class)
```

```
arandi(mbind2$cl, tru.class)
arandi(relab2$cl, tru.class)
```

`arandi` *(Adjusted) Rand Index for Clusterings*

Description

Computes the adjusted or unadjusted Rand index between two clusterings/partitions of the same objects.

Usage

```
arandi(cl1, cl2, adjust = TRUE)
```

Arguments

`cl1,cl2`	vectors of cluster memberships (need to have the same lengths).
`adjust`	logical. Should index be adjusted? Defaults to TRUE.

Details

The Rand index is based on how often the two clusterings agree in the treatment of pairs of observations, where agreement means that two observations are in/not in the same cluster in both clusterings.

The adjusted Rand index adjusts for the expected number of chance agreements.

Formulas of Hubert and Arabie (1985) are used for the computation.

References

Hubert, L. and Arabie, P. (1985): Comparing partitions. *Journal of Classification*, **2**, 193–218.

See Also

`vi.dist`

Examples

```
cl1 <- sample(1:3,10,replace=TRUE)
cl2 <- c(cl1[1:5], sample(1:3,5,replace=TRUE))
arandi(cl1,cl2)
arandi(cl1,cl2,adjust=FALSE)
```

`cls.draw1.5` *Sample of Clusterings from Posterior Distribution of Bayesian Cluster Model*

Description

Output of a Dirichlet process mixture model with normal components fitted to the data set `Ysim1.5`. True clusters are given by `rep(1:8,each =50)`.

Usage

```
data(cls.draw1.5)
```

Format

matrix with 500 rows and 400 columns. Each row contains a clustering of the 400 observations.

`cls.draw2` *Sample of Clusterings from Posterior Distribution of Bayesian Cluster Model*

Description

Output of a Dirichlet process mixture model with normal components fitted to the data set `Ysim2`. True clusters are given by `rep(1:8,each =50)`.

Usage

```
data(cls.draw2)
```

Format

matrix with 500 rows and 400 columns. Each row contains a clustering of the 400 observations.

`comp.psm` *Estimate Posterior Similarity Matrix*

Description

For a sample of clusterings of the same objects the proportion of clusterings in which observation i and j are together in a cluster is computed and a matrix containing all proportions is given out.

Usage

```
comp.psm(cls)
```

Arguments

`cls`	a matrix in which every row corresponds to a clustering of the `ncol(cls)` objects

Details

In Bayesian cluster analysis the posterior similarity matrix is a matrix whose entry $[i, j]$ contains the posterior probability that observation i and j are together in a cluster. It is estimated by the proportion of a posteriori clusterings in which i and j cluster together.

Value

a symmetric `ncol(cls)*ncol(cls)` matrix

See Also

`cltoSim`

Examples

```
(cls <- rbind(c(1,1,2,2),c(1,1,2,2),c(1,2,2,2),c(2,2,1,1)))
comp.psm(cls)
```

`maxpear` *Maximize/Compute Posterior Expected Adjusted Rand Index*

Description

Based on a posterior similarity matrix of a sample of clusterings `maxpear` finds the clustering that maximizes the posterior expected Rand adjusted index (PEAR) with the true clustering, while `pear` computes PEAR for several provided clusterings.

Usage

```
maxpear(psm, cls.draw = NULL, method = c("avg", "comp", "draws",
         "all"), max.k = NULL)

pear(cls,psm)
```

Arguments

`psm`	a posterior similarity matrix, usually obtained from a call to `comp.psm`.
`cls, cls.draw`	a matrix in which every row corresponds to a clustering of the `ncol(cls)` objects. `cls.draw` refers to the clusterings that have been used to compute `psm`, `cls.draw` has to be provided if `method="draw"` or `"all"`.
`method`	the maximization method used. Should be one of `"avg"`, `"comp"`, `"draws"` or `"all"`. The default is `"avg"`.
`max.k`	integer, if `method="avg"` or `"comp"` the maximum number of clusters up to which the hierarchical clustering is cut. Defaults to `ceiling(nrow(psm)/8)`.

Details

For `method="avg"` and `"comp"` `1-psm` is used as a distance matrix for hierarchical clustering with average/complete linkage. The hierarchical clustering is cut for the cluster sizes `1:max.k` and PEAR computed for these clusterings.
Method `"draws"` simply computes PEAR for each row of `cls.draw` and takes the maximum.
If `method="all"` all maximization methods are applied.

Value

`cl`	clustering with maximal value of PEAR. If `method="all"` a matrix containing the clustering with the highest value of PEAR over all methods in the first row and the clusterings of the individual methods in the next rows.
`value`	value of PEAR. A vector corresponding to the rows of `cl` if `method="all"`.
`method`	the maximization method used.

See Also

`comp.psm` for computing posterior similarity matrix, `minbinder`, `medv`, `relabel` for other possibilities for processing a sample of clusterings.

Examples

```
data(cls.draw1.5)
# sample of 500 clusterings from a Bayesian cluster model
tru.class <- rep(1:8,each=50)
# the true grouping of the observations
psm1.5 <- comp.psm(cls.draw1.5)
mpear1.5 <- maxpear(psm1.5)
table(mpear1.5$cl, tru.class)

# Does hierarchical clustering with Ward's method lead
# to a better value of PEAR?
hclust.ward <- hclust(as.dist(1-psm1.5), method="ward")
cls.ward <-
t(apply(matrix(1:20),1, function(k) cutree(hclust.ward,k=k)))
ward1.5 <- pear(cls.ward, psm1.5)
max(ward1.5) > mpear1.5$value
```

`medv` *Clustering Method of Medvedovic*

Description

Based on a posterior similarity matrix of a sample of clusterings `medv` obtains a clustering by using `1-psm` as distance matrix for hierarchical clustering with complete linkage. The dendrogram is cut at a value `h` close to 1.

Usage

```
medv(psm, h=0.99)
```

Arguments

`psm`	a posterior similarity matrix, usually obtained from a call to `comp.psm`.
`h`	The height at which the dendrogram is cut.

Value

vector of cluster memberships.

References

Medvedovic, M. Yeung, K. and Bumgarner, R. (2004) Bayesian mixture model based clustering of replicated microarray data, *Bioinformatics*, **20**, 1222-1232.

See Also

`comp.psm` for computing posterior similarity matrix, `maxpear`, `minbinder`, `relabel` for other possibilities for processing a sample of clusterings.

Examples

```
data(cls.draw1.5)
# sample of 500 clusterings from a Bayesian cluster model
tru.class <- rep(1:8,each=50)
# the true grouping of the observations
psm1.5 <- comp.psm(cls.draw1.5)
medv1.5 <- medv(psm1.5)
table(medv1.5, tru.class)
```

`minbinder` *Minimize/Compute Posterior Expectation of Binders Loss Function*

Description

Based on a posterior similarity matrix of a sample of clusterings `minbinder` finds the clustering that minimizes the posterior expectation of Binders loss function, while `binder` computes the posterior expected loss for several provided clusterings.

Usage

```
minbinder(psm, cls.draw = NULL, method = c("avg", "comp", "draws",
          "laugreen","all"), max.k = NULL, include.lg = FALSE,
          start.cl = NULL, tol = 0.001)

binder(cls,psm)

laugreen(psm, start.cl, tol=0.001)
```

Arguments

`psm`	a posterior similarity matrix, usually obtained from a call to `comp.psm`.
`cls, cls.draw`	a matrix in which every row corresponds to a clustering of the `ncol(cls)` objects. `cls.draw` refers to the clusterings that have been used to compute `psm`, `cls.draw` has to be provided if `method="draw"` or `"all"`.
`method`	the maximization method used. Should be one of `"avg"`, `"comp"`, `"draws"`, `"laugreen"` or `"all"`. The default is `"avg"`.
`max.k`	integer, if `method="avg"` or `"comp"` the maximum number of clusters up to which the hierarchical clustering is cut. Defaults to `ceiling(nrow(psm)/4)`.
`include.lg`	logical, should method `"laugreen"` be included when `method="all"`? Defaults to FALSE.
`start.cl`	clustering used as starting point for `method="laugreen"`. If `NULL` `start.cl= 1:nrow(psm)` is used.
`tol`	convergence tolerance for `method="laugreen"`.

Details

The posterior expected loss is the sum of the absolute differences of the indicator function of observation i and j clustering together and the posterior probability that they are in one cluster.

For `method="avg"` and `"comp"` `1-psm` is used as a distance matrix for hierarchical clustering with average/complete linkage. The hierarchical clustering is cut for the cluster sizes `1:max.k` and the posterior expected loss is computed for these clusterings.
Method `"draws"` simply computes the posterior expected loss for each row of `cls.draw` and takes the minimum.
Method `"laugreen"` implements the algorithm of Lau and Green (2007), which

is based on binary integer programming. Since the method can take some time to converge it is only used if explicitly demanded with `method="laugreen"` or `method="all"` *and* `include.lg=TRUE`. If `method="all"` all minimization methods except `"laugreen"` are applied.

Value

`cl`	clustering with minimal value of expected loss. If `method="all"` a matrix containing the clustering with the smallest value of the expected loss over all methods in the first row and the clusterings of the individual methods in the next rows.
`value`	value of posterior expected loss. A vector corresponding to the rows of `cl` if `method="all"`.
`method`	the maximization method used.
`iter.lg`	if `method="laugreen"` the number of iterations the method needed to converge.

References

Binder, D.A. (1978) Bayesian cluster analysis, *Biometrika* **65**, 31–38.

Fritsch, A. and Ickstadt, K. (2009) An improved criterion for clustering based on the posterior similarity matrix, *Bayesian Analysis*, **4**, 367–392.

Lau, J.W. and Green, P.J. (2007) Bayesian model based clustering procedures, *Journal of Computational and Graphical Statistics* **16**, 526–558.

See Also

`comp.psm` for computing posterior similarity matrix, `maxpear`, `medv`, `relabel` for other possibilities for processing a sample of clusterings. `lp` for the linear programming.

Examples

```
data(cls.draw2)
# sample of 500 clusterings from a Bayesian cluster model
tru.class <- rep(1:8,each=50)
# the true grouping of the observations
psm2 <- comp.psm(cls.draw2)
mbind2 <- minbinder(psm2)
table(mbind2$cl, tru.class)

# Does hierarchical clustering with Ward's method lead
```

```
# to a lower value of Binders loss?
hclust.ward <- hclust(as.dist(1-psm2), method="ward")
cls.ward <-
t(apply(matrix(1:20),1, function(k) cutree(hclust.ward,k=k)))
ward2 <- binder(cls.ward, psm2)
min(ward2) < mbind2$value

# Method laugreen is applied to 40 randomly selected observations
ind <- sample(1:400, 40)
mbind.lg <- minbinder(psm2[ind, ind],cls.draw2[,ind], method="all"
                        ,include.lg=TRUE)
mbind.lg$value
```

`norm.label` *Norm Labelling of a Clustering*

Description

Cluster labels of a clusterings are replaced by `1:length(table(cl))`.

Usage

```
norm.label(cl)
```

Arguments

`cl`	vector of cluster memberships

Value

the clustering with normed labels.

See Also

`relabel` for labelling a sample of clusterings the same way

Examples

```
(cl <- sample(c(13,12,34), 13, replace=TRUE))
norm.label(cl)

(cl <- sample(c("a","b","f31"), 13, replace=TRUE))
norm.label(cl)
```

`relabel` *Stephens' Relabeling Algorithm for Clusterings*

Description

For a sample of clusterings in which corresponding clusters have different labels the algorithm attempts to bring the clusterings to a unique labelling.

Usage

```
relabel(cls, print.loss = TRUE)
```

Arguments

`cls`	a matrix in which every row corresponds to a clustering of the `ncol(cls)` objects.
`print.loss`	logical, should current value of loss function be printed after each iteration? Defaults to TRUE.

Details

The algorithm minimizes the loss function

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{j=1}^{K} -\log \hat{p}_{ij} \cdot I_{\{z_i^{(m)}=j\}}$$

over the M clusterings, n observations and K clusters, where $\hat{p}_{ij}$ is the estimated probability that observation i belongs to cluster j and $z_i^{(m)}$ indicates to which cluster observation i belongs in clustering m. $I_{\{.\}}$ is an indicator function.

Minimization is achieved by iterating the estimation of $\hat{p}_{ij}$ over all clusterings and the minimization of the loss function in each clustering by permuting the cluster labels. The latter is done by linear programming.

Value

`cls`	the input `cls` with unified labelling.
`P`	an $n \times K$ matrix, where entry $[i, j]$ contains the estimated probability that observation i belongs to cluster j.
`loss.val`	value of the loss function.
`cl`	vector of cluster memberships that have the highest probabilities $\hat{p}_{ij}$.

Warning

The algorithm assumes that the number of clusters K is fixed. If this is not the case K is taken to be the most common number of clusters. Clusterings with other numbers of clusters are discarded and a warning is issued.

Note

The implementation is a variant of the algorithm of Stephens which is originally applied to draws of parameters for each observation, not to cluster labels.

References

Stephens, M. (2000) Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society Series B*, **62**, 795–809.

See Also

`lp.transport` for the linear programming, `maxpear`, `minbinder`, `medv` for other possibilities of processing a sample of clusterings.

Examples

```
(cls <- rbind(c(1,1,2,2),c(1,1,2,2),c(1,2,2,2),c(2,2,1,1)))
# group 2 in clustering 4 corresponds to group 1 in clustering 1-3.
cls.relab <- relabel(cls)
cls.relab$cls
```

`vi.dist` *Variation of Information Distance for Clusterings*

Description

Computes the 'variation of information' distance of Meila (2007) between two clusterings/partitions of the same objects.

Usage

```
vi.dist(cl1, cl2, parts = FALSE, base = 2)
```

Arguments

`cl1,cl2`	vectors of cluster memberships (need to have the same lengths).
`parts`	logical; should the two conditional entropies also be returned?
`base`	base of logarithm used for computation of entropy and mutual information.

Details

The variation of information distance is the sum of the two conditional entropies of one clustering given the other. For details see Meila (2007).

Value

The VI distance. If `parts=TRUE` the two conditional entropies are appended.

References

Meila, M. (2007) Comparing Clusterings - an Information Based Distance. *Journal of Multivariate Analysis*, **98**, 873 – 895.

See Also

`arandi`

Examples

```
cl1 <- sample(1:3,10,replace=TRUE)
cl2 <- c(cl1[1:5], sample(1:3,5,replace=TRUE))
vi.dist(cl1,cl2)
vi.dist(cl1,cl2, parts=TRUE)
```

`Ysim1.5` *Simulated 3-dimensional Normal Data Containing 8 Clusters*

Description

Cluster means are given by the 8 possible values of $(\pm 1.5, \pm 1.5, \pm 1.5)$ to which standard normal noise was added. True clusters are given by `rep(1:8,each =50)`.

Usage

```
data(Ysim1.5)
```

Format

matrix with 400 rows and 3 columns.

Source

Simulated by

```
1.5 * matrix(c(rep(c(1,1,1),50), rep(c(1,1,-1),50), rep(c(1,-
1,1),50), rep(c(-1,1,1),50), rep(c(1,-1,-1),50), rep(c(-1,1,-
1),50), rep(c(-1,-1,1),50), rep(c(-1,-1,-1),50)), byrow=TRUE,
ncol=3) + matrix(rnorm( 400*3),ncol=3)
```

`Ysim2` *Simulated 3-dimensional Normal Data Containing 8 Clusters*

Description

Cluster means are given by the 8 possible values of $(\pm 2, \pm 2, \pm 2)$ to which standard normal noise was added. True clusters are given by `rep(1:8,each =50)`.

Usage

```
data(Ysim2)
```

Format

matrix with 400 rows and 3 columns.

Source

Simulated by

```
2 * matrix(c(rep(c(1,1,1),50), rep(c(1,1,-1),50), rep(c(1,-
1,1),50), rep(c(-1,1,1),50), rep(c(1,-1,-1),50), rep(c(-1,1,-
1),50), rep(c(-1,-1,1),50), rep(c(-1,-1,-1),50)), byrow=TRUE,
ncol=3) + matrix(rnorm( 400*3),ncol=3)
```

Bibliography

Akaike, H. (1974). "A New Look at Statistical Model Identification." *IEEE Transactions on Automatic Control*, 19: 716–723.

Aldous, D. J. (1985). *Exchangeability and Related Topics*. Berlin: Springer.

Anderson, E. (1935). "The Irises of the Gaspé Peninsula." *Bulletin of the American Iris Scociety*, 59: 2–5.

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *Annals of Statistics*, 2: 1152–1174.

Bansal, N., Blum, A., and Chawla, S. (2004). "Correlation Clustering." *Machine Learning*, 56: 89–113.

Barron, A. (1988). "The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions." Technical report, 1988-7, Department of Statistics, University of Illinois.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). "The Consistency of Posterior Distributions in Nonparametric Problems." *Annals of Statistics*, 27: 536–561.

Bell, E. T. (1934). "Exponential Numbers." *American Mathematical Monthly*, 41: 411–419.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis." *Statistics and Computing*, 7: 1–10.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Binder, D. A. (1978). "Bayesian Cluster Analysis." *Biometrika*, 65: 31–38.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions via Pólya Urn Schemes." *Annals of Statistics*, 1: 353–355.

Bornkamp, B., Fritsch, A., Kuß, O., and Ickstadt, K. (2008). "Penalty Specialists Among Goalkeepers - A Nonparametric Bayesian Analysis of 44 Years of the German Bundesliga." In Schipp, B. and Krämer, W. (eds.), *Statistical Inference, Econometric Analysis and Matrix Algebra. Festschrift in Honour of Götz Trenkler.*, 63–76. Heidelberg: Physica-Verlag.

Bornkamp, B., Ickstadt, K., and Dunson, D. B. (2010). "Stochastically Ordered Multiple Regression." *Biostatistics*, to appear.

Bush, C. A. and MacEachern, S. N. (1996). "A Semiparametric Bayesian Model for Randomised Block Designs." *Biometrika*, 83: 275–285.

Chen, R. and Liu, J. S. (1996). "Predictive Updating Methods with Application to Bayesian Classification." *Journal of the Royal Statistical Society, Series B*, 58: 397–415.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Cox, T. F. and Cox, M. A. (2001). *Multidimensional Scaling*. Boca Raton: Chapman&Hall, 2nd edition.

Dahl, D. B. (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model." In Do, K. A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 201–218. Cambridge University Press.

— (2007). "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, under revision.

Dahl, D. B. and Newton, M. A. (2007). "Multiple Hypothesis Testing by Clustering Treatment Effects." *Journal of the American Statistical Association*, 102: 517–526.

Dalal, S. R. and Hall, W. J. (1983). "Approximating Priors by Mixtures of Natural Conjugate Priors." *Journal of the Royal Statistical Society, Series B*, 45: 278–286.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B*, 39: 1–38.

Diaconis, P. and Ylvisaker, D. (1985). "Quantifying Prior Opinion." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 2*, 133–156. Elsevier Science Publishers B.V.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, 97: 77–87.

Dunson, D. B. (2010). "Nonparametric Bayes Applications to Biostatistics." In Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, to appear. Cambridge University Press.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). "Cluster Analysis and Display of Genome-Wide Expression Patterns." *Proceedings of the National Academy of Sciences (USA)*, 95: 14863–14868.

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90: 577–588.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis*. London: Arnold, 4th edition.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics*, 1: 209–230.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, 7: 179–188.

Fowlkes, E. B. and Mallows, C. L. (1983). "A Method for Comparing Two Hierarchical Clusterings." *Journal of the American Statistical Association*, 78: 553–569.

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis and Density Estimation." *Journal of the American Statistical Association*, 97: 611–631.

— (2007). "Bayesian Reguralization for Normal Mixture Estimation and Model-Based Clustering." *Journal of Classification*, 24: 155–181.

Freedman, D. A. (1963). "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case." *Annals of Mathematical Statistics*, 34: 1386–1403.

Friedman, H. P. and Rubin, J. (1967). "On Some Invariant Criteria for Grouping Data." *Journal of the American Statistical Association*, 63: 1159–1178.

Fritsch, A. and Ickstadt, K. (2009). "Improved Criteria for Clustering Based on the Posterior Similarity Matrix." *Bayesian Analysis*, 4: 367–392.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Berlin: Springer.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing." *Journal of the American Statistical Association*, 100: 1021–1035.

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches for Calculating Marginal Densities." *Journal of the American Statistical Association*, 85: 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman&Hall, 2nd edition.

Gene Ontology Consortium (2000). "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics*, 25: 25–29.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, 5: R80.

Ghosal, S., Ghosh, J., and Ramamoorthi, R. (1999). "Posterior Consistency of Dirichlet Mixtures in Density Estimation." *Annals of Statistics*, 27: 143–158.

Gibbs, A. and Su, F. E. (2002). "On Choosing and Bounding Probability Metrics." *International Statistical Review*, 70: 419–435.

Goder, A. and Filkov, V. (2008). "Consensus Clustering Algorithms: Comparison and Refinement." In *ALENEX 2008. Proceedings in Algorithm Engineering and Experiments*, 109–117. SIAM.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, 286: 531–537.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1989). *Concrete Mathematics: Foundation for Computer Science*. Reading: Addison-Wesley.

Green, P. J. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, 82: 711–732.

Green, P. J. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28: 355–375.

Griffin, J. E. and Steel, M. F. J. (2006). "Order-Based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 101: 179–194.

Hartigan, J. A. (1990). "Partition Models." *Communications in Statistics, Part A - Theory and Methods*, 19: 2745–2756.

Hathaway, R. J. (1985). "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions." *Annals of Statistics*, 13: 795–800.

Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G. (eds.) (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. New York: Wiley, 2nd edition.

Hubert, L. and Arabie, P. (1985). "Comparing Partitions." *Journal of Classification*, 2: 193–218.

Hurn, M., Justel, A., and Robert, C. P. (2003). "Estimating Mixtures of Regressions." *Journal of Computational and Graphical Statistics*, 12: 55–79.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network." *Science*, 292: 929–934.

Ishwaran, H. (2000). "Inference for the Random Effects in Bayesian Generalized Linear Mixed Models." *ASA Proceedings of the Bayesian Statistical Science Section*, 1–10.

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173.

Jara, A., Hanson, T., Quintana, F. A., Müller, P., and Rosner, G. L. (2009).

"DPpackage: Bayesian Nonparametric and Semiparametric Analysis." R package: 1.0–7.

Jensen, S. T. and Liu, J. S. (2008). "Bayesian Clustering of Transcription Factor Binding Motifs." *Journal of the American Statistical Association*, 103: 188–200.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Volume 2*. New York: Wiley, 2nd edition.

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: Wiley.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). "Variable Selection in Clustering via Dirichlet Process Mixture Models." *Biometrika*, 93: 877–893.

Lance, G. N. and Williams, W. T. (1966). "A General Theory for Classificatory Sorting Strategies: 1. Hierarchical Systems." *Computer Journal*, 9: 60–64.

Lau, J. W. and Green, P. J. (2007). "Bayesian Model-Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558.

Leininger, W. and Ockenfels, A. (2008). "The Penalty-Duel and Institutional Design: Is there a Neeskens-Effect?" In Anderson, P., Ayton, P., and Schmidt, C. (eds.), *Myths and Facts about Football: The Economics*

and Psychology of the World´s Greatest Sport, 73–93. Cambridge: Cambridge Scholar Press.

Leisch, F. (2004). "Exploring the Structure of Mixture Model Components." In Antoch, J. (ed.), *COMPSTAT 2004. Proceedings in Computational Statistics*, 1405–1412. Heidelberg: Physica-Verlag/Springer.

Li, J., Ray, S., and Lindsay, B. G. (2007). "A Nonparametric Statistical Approach to Clustering via Mode Identification." *Journal of Machine Learning Research*, 8: 1687–1723.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the Reinforcement in Bayesian Non-Parametric Mixture Models." *Journal of the Royal Statistical Society, Series B*, 69: 715–740.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2006). "Loss Function Based Ranking in Two-Stage Hierarchical Models." *Bayesian Analysis*, 1: 915–946.

MacEachern, S. N. (2000). "Dependent Dirichlet Processes." Technical report, Department of Statistics, Ohio State University.

MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations." In LeCam, L. and Neyman, J. (eds.), *5th Berkeley Symposium on Mathematical Statistics and Probability*.

Marron, J. S. and Wand, M. P. (1992). "Exact Mean Integrated Squared Error." *Annals of Statistics*, 20: 712–736.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). "Bayesian Mixture Model Based Clustering of Replicated Microarray Data." *Bioinformatics*, 20: 1222–1232.

Meilă, M. (2007). "Comparing Clusterings – An Information Based Distance." *Journal of Multivariate Analysis*, 98: 873–895.

Meilă, M. and Heckerman, D. (2001). "An Experimental Comparison of Model-Based Clustering Methods." *Machine Learning*, 42: 9–29.

Milligan, G. W. and Cooper, M. C. (1986). "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis." *Multivariate Behavioral Research*, 21: 441–458.

Mirkin, B. (1996). *Mathematical Classification and Clustering*. Dordrecht: Kluwer Academic Press.

— (2005). *Clustering for Data Mining - A Data Recovery Approach*. Boca Raton: Chapman&Hall.

Neal, R. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9: 249–265.

Newton, M. A., Czado, C., and Chappell, R. (1996). "Bayesian Inference for Semiparametric Bayesian Binary Regression." *Journal of the American Statistical Association*, 91: 142–153.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). "Flexible Random-Effects Models Using Bayesian Semi-Parametric Models: Applications to Institutional Comparisons." *Statistics in Medicine*, 26: 2088–2112.

Ongaro, A. and Cattaneo, C. (2004). "Discrete Random Probability Measures: A General Framework for Nonparametric Bayesian Inference." *Statistics & Probability Letters*, 67: 33–45.

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models." *Biometrika*, 95: 169–186.

Pearson, K. (1894). "Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London A*, 185: 71–110.

Pitman, J. (2006). *Combinatorial Stochastic Processes - Ecole d'Eté de Probabilités de Saint-Flour 2002*. Berlin: Springer.

Pitman, J. and Yor, M. (1997). "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator." *Annals of Probability*, 25: 855–900.

Podwojski, K., Fritsch, A., Chamrad, D. C., Paul, W., Sitek, B., Stühler, K., Mutzel, P., Stephan, C., Meyer, H. E., Urfer, W., Ickstadt, K., and Rahnenführer, J. (2009). "Retention Time Alignment Algorithms for LC/MS Data must Consider Nonlinear Shifts." *Bioinformatics*, 25: 758–764.

Qin, Z. S. (2006). "Clustering Microarray Gene Expression Data Using Weighted Chinese Restaurant Process." *Bioinformatics*, 22: 1988–1997.

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian Clustering and Product Partition Models." *Journal of the Royal Statistical Society, Series B*, 65: 557–574.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66: 846–850.

Ray, S. and Lindsay, B. G. (2005). "The Topography of Multivariate Normal Mixtures." *Annals of Statistics*, 33: 2042–2065.

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society, Series B*, 59: 731–792.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichlet Process (with Discussion)." *Journal of the American Statistical Association*, 103: 1131–1154.

Roeder, K. and Wasserman, L. (1997). "Practical Bayesian Density Estimation Using Mixtures of Normals." *Journal of the American Statistical Association*, 92: 894–902.

Rota, G.-C. (1964). "The Number of Partitions of a Set." *American Mathematical Monthly*, 71: 498–504.

Schwartz, L. (1965). "On Bayes Procedures." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4: 10–26.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, 6: 461–464.

Scott, A. J. and Symons, M. J. (1971). "Clustering Methods Based on Likelihood Ratio Criteria." *Biometrics*, 27: 387–397.

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4: 639–650.

Sokal, R. R. and Michener, C. D. (1958). "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin*, 38: 1409–1438.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B*, 64: 583–639.

Steinley, D. (2004). "Properties of the Hubert-Arabie Adjusted Rand Index." *Psychological Methods*, 9: 386–396.

Stephens, M. (1997). "Bayesian Methods for Mixtures of Normal Distributions." Ph.D. thesis, Department of Statistics, University of Oxford.

— (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society, Series B*, 62: 795–809.

Symons, M. J. (1981). "Clustering Criteria and Multivariate Normal Mixtures." *Biometrics*, 37: 35–43.

Tadesse, M. G., Sha, N., Kim, S., and Vannucci, M. (2006). "Identification of Biomarkers in Classification and Clustering of High-Throughput Data." In Do, K. A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 97–115. Cambridge University Press.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). "Bayesian Variable Selection in Clustering High-Dimensional Data." *Journal of the American Statistical Association*, 100: 602–617.

Teh, Y. W., Jordan, M., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101: 1566–1581.

Teicher, H. (1963). "Identifiability of Finite Mixtures." *Annals of Mathematical Statistics*, 34: 1265–1269.

Treppmann, T. (2010). "Vergleich externer Kriterien zur Cluster-Validierung." Bachelor's thesis, Department of Statistics, Technische Universität Dortmund.

Verdinelli, I. and Wasserman, L. (1991). "Bayesian Analysis of Outlier Problems Using the Gibbs Sampler." *Statistics and Computing*, 1: 105–117.

Walker, S. G. (2004). "New Approaches to Bayesian Consistency." *Annals of Statistics*, 32: 2028–2043.

Ward, H. J., Jr. (1963). "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, 58: 236–244.

Wilson, S. R. (1982). "Sound and Exploratory Data Analysis." In *COMPSTAT 1982. Proceedings in Computational Statistics*, 447–450. Vienna: Physica-Verlag.

Wishart, D. (1969). "An Algorithm for Hierarchical Classification." *Biometrics*, 25: 165–170.

Wu, Y. and Ghosal, S. (2008). "Kullback Leibler Property of Kernel Mixture Priors in Bayesian Density Estimation." *Electronic Journal of Statistics*, 2: 298–331.

Yakowitz, S. J. and Spragins, J. D. (1968). "On the Identifiability of Finite Mixtures." *Annals of Mathematical Statistics*, 39: 209–214.